Learning Registered Point Processes from Idiosyncratic Observations

Hongteng Xu [1 2] **Lawrence Carin** [1] **Hongyuan Zha** [3]

Abstract

A parametric point process model is developed, with modeling based on the assumption that sequential observations often share latent phenomena, while also possessing idiosyncratic effects. An alternating optimization method is proposed to learn a "registered" point process that accounts for shared structure, as well as "warping" functions that characterize idiosyncratic aspects of each observed sequence. Under reasonable constraints, in each iteration we update the sample-specific warping functions by solving a set of constrained nonlinear programming problems in parallel, and update the model by maximum likelihood estimation. The justifiability, complexity and robustness of the proposed method are investigated in detail, and the influence of sequence stitching on the learning results is discussed empirically. Experiments on both synthetic and real-world data demonstrate that the method yields explainable point process models, achieving encouraging results compared to state-of-the-art methods.

1. Introduction

The behavior of real-world entities often may be recorded as event sequences; for example, interactions of participants in a social network, the admissions of patients, and the job-hopping behavior of employees. In practice, these behaviors are under the control of complicated mechanisms, which can be captured approximately by an appropriate parametric temporal point process model. While the observed event sequences associated with a given process ($e.g.$, disease) may share common ("standard") attributes, there are often subject-specific factors that may impact the observed data. For example, the admission records of different patients are always personalized: even if the patients suffer from the same disease, they may spend unequal time on recovery because their medications, history and environmental conditions may be distinct. Another typical example is the job-hopping behavior of employees. The employees in the same company often make very different career plans depending on their age, family situation and unobserved status of the job market.

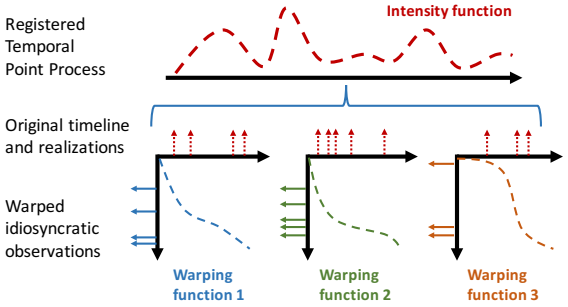


Figure 1. The illustration of concepts in our work. The dotted parts (parametric point process model, unwarped realizations and warping functions) are what we aim to learn.

The examples above reveal that event sequences that share an underlying temporal point process linked to a given phenomenon of interest may be personalized by hidden idiosyncratic factors, yielding a subject-specific "warping" along timeline, as shown in Fig. 1. The characteristics of such data often have a negative influence on the learning of the target point process, $i.e.$, increase the uncertainty of the model. The complexity of models can be increased to fit the personalized observations well, $e.g.$, the locally-stationary point processes in (Roueff et al., 2016; Mammen, 2017; Xu et al., 2017a). However, from the viewpoint of model registration, it is desirable to separate the essential mechanism of the model and idiosyncratic aspects of the data, such that the final model is "registered" and characterizes the shared phenomena, while also inferring what is sample-specific.

Learning registered point processes from idiosyncratic observations is a challenging problem, requiring one to jointly learn a shared point process model and a set of sample-specific warping functions corresponding to observed event sequences. To solve this problem, we propose a novel and effective learning method based on alternating optimization. Specifically, in each iteration we first apply the inverse of estimated warping functions ($i.e.$, unwarping functions) to

[1]Department of ECE, Duke University, Durham, NC, USA [2]InfiniaML Inc., Durham, NC, USA [3]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Hongteng Xu <hongteng.xu@duke.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

unwarp observed event sequences and learn the parameter of a registered point process by maximum likelihood estimation; we then update the warping functions of event sequences, based on the estimation of registered point process. The new functions are applied to update the model for the next iteration. In particular, we approximate the warping/unwarping functions of event sequences by piecewise linear models, and learn their parameters by solving a set of constrained nonlinear programming problems in parallel.

We analyze the justification and the complexity of our method in detail. The meaning of the regularizers and constraints used in our method and their effects are also investigated. Furthermore, we consider to improve learning results by stitching warped sequences randomly and learning from the stitched sequences, and verify the feasibility of this data processing strategy empirically. Experimental results show that the proposed method outperforms its competitors on both synthetic and real-world data.

2. Proposed Model

Denote a temporal point process as N. Its event sequence consists of multiple events $\{(t_i, c_i)\}_{i=1}^{I}$ with time stamps $t_i \in [0, T]$ and event types $c_i \in \mathcal{C} = \{1, ..., C\}$, which can be represented as $\{N_c(t)\}_{c=1}^{C}$, where $N_c(t)$ is the number of type-c events occurring at or before time t. A temporal point process can be characterized by its intensity functions $\{\lambda_c(t)\}_{c=1}^{C}$, where $\lambda_c(t) = \mathbb{E}[dN_c(t)|\mathcal{H}_t^{\mathcal{C}}]/dt$ and $\mathcal{H}_t^{\mathcal{C}} = \{(t_i, c_i)|t_i < t, c_i \in \mathcal{C}\}$ collects historical events before time t. Each $\lambda_c(t)$ represents the expected instantaneous rate of the type-c event at time t, which can be parametrized by a parameter θ. In this work, we assume that:

1. Exponential-like intensity: each $\lambda(t)$ is an exponential-like function $\sum_{j=1}^{J} \exp_{t_j}(f_j(t; \theta, \mathcal{H}_t^{\mathcal{C}}))$, where $J \geq 1$, f_j's are linear functions of time, which are related to θ and historical observations. $\exp_{t_j}(f_j(t)) = \exp(f_j(t))$ *if* $t \geq t_j$*, otherwise, it equals to* 0.

Note that many important point processes, $e.g.$, the Hawkes process (Hawkes & Oakes, 1974) and the self-correcting process (Isham & Westcott, 1979) satisfy this assumption, as shown in the Appendix. The sequences of a parametric point process N_θ may be warped in $[0, T]$ by a set of continuous and invertible warping functions. Denote the sequences and the corresponding warping functions as $\{S_m\}_{m=1}^{M}$ and $\{W_m\}_{m=1}^{M}$, respectively. Each $S_m = \{(t_i^m, c_i^m)\}_{i=1}^{I_m}$ contains I_m events, whose time stamps are deformed from a "standard" timeline under the corresponding warping function $W_m : [0, T] \mapsto [0, T]$. Accordingly, the unwarping functions can be denoted as $\{W_m^{-1}\}_{m=1}^{M}$. We assume that for $m = 1, ..., M$

2. Unbiasedness: $\mathbb{E}[W_m^{-1}(t)] = t$ *on* $[0, T]$.
3. Regularity: $W_m^{-1}(t)$ *is monotone increasing on* $[0, T]$.

Taking the warping functions into account, the likelihood of an (unobserved) unwarped sequence can be formulated based on the intensity functions (Daley & Vere-Jones, 2007):

$$\mathcal{L}(\theta; W_m^{-1}(S_m)) = \frac{\prod_{i=1}^{I_m} \lambda_{c_i^m}(W_m^{-1}(t_i^m))}{\exp\left(\sum_{c=1}^{C} \int_0^T \lambda_c(W_m^{-1}(s))ds\right)}, \quad (1)$$

where $W_m^{-1}(S_m)$ represents the unwarped event sequence, $i.e.$, $W_m^{-1}(S_m) = \{(W_m^{-1}(t_i^m), c_i^m)\}_{i=1}^{I_m}$.

The warped data caused by idiosyncratic effects generally do harm to the maximum likelihood estimation of the target point process, except some trivial cases (See Appendix 9.2):

Proposition 2.1. *For a temporal point process N_θ satisfying the assumption 3, $\hat{\theta}^*$ and $\hat{\theta}$ denote its maximum likelihood estimation based on original data and that based on warped data, respectively. Then $\hat{\theta}^* = \hat{\theta}$ if and only if 1) the warping functions are translations; or 2) N_θ is a homogeneous Poisson process.*

The problem is that given the warped observations $\{S_m\}_{m=1}^{M}$, we seek to learn a "registered" model θ, as well as sample-specific warping functions $\{W_m\}_{m=1}^{M}$ (or equivalently, the unwarping functions $\{W_m^{-1}\}_{m=1}^{M}$) .

3. Learning Registered Point Processes

3.1. Maximizing the likelihood

We develop a learning method based on maximum likelihood estimation (MLE). Considering the assumptions of unwarping functions and the likelihood in (1), we can formulate the optimization problem as

$$\begin{aligned} \min_{\theta, \{W_m\}} \quad & -\sum\nolimits_m \log \mathcal{L}(\theta; W_m^{-1}(S_m)) + \gamma \mathcal{R}(\{W_m^{-1}\}) \\ s.t.\ & 1)\ W_m^{-1}(0) = 0,\ W_m^{-1}(T) = T, \text{ and} \\ & 2)\ {W_m^{-1}}'(t) > 0 \text{ for } m = 1, ..., M, \end{aligned} \quad (2)$$

where ${W_m^{-1}}'(t) = \frac{dW_m'}{dt}$ is the derivative of unwarping function. In our objective function, the first term represents the negative log-likelihood of unwarped event sequences while the second term represents the regularizer imposed on unwarping functions. For each unwarping function, the first constraint corresponds to its range and the second constraint makes it obey the regularity assumption. Furthermore, according to the unbiasedness assumption, we apply the following regularizer:

$$\mathcal{R}(\{W_m^{-1}\}) = \int_0^T \left| \frac{1}{M} \sum\nolimits_{m=1}^{M} W_m^{-1}(s) - s \right|^2 ds. \quad (3)$$

The optimization problem in (2) is non-convex and has a large number of unknown variables. Solving it directly is

intractable. Fortunately, for the parametric point processes with exponential-like intensity functions, we can design an effective alternating optimization method to solve the problem iteratively, after parameterizing the warping functions as piecewise linear functions. In each iteration, we first maximize the likelihood of the unwarped sequences based on the estimation of warping functions, and then optimize the warping functions based on the estimated model.

Specifically, in the k-th iteration, given the warping functions estimated in the previous iteration, *i.e.*, $\{W_m^{k-1}\}_{m=1}^M$, we learn the target point process by

$$\theta^k = \arg\min_\theta - \sum_{m=1}^M \log \mathcal{L}(\theta; (W_m^{k-1})^{-1}(S_m)). \quad (4)$$

Focusing on different point processes, we can apply various optimization methods to solve this problem. For example, learning Hawkes processes can be achieved in the framework of expectation-maximization (EM) (Lewis & Mohler, 2011; Zhou et al., 2013), which is equivalent to a projected-gradient-ascent algorithm. For other kinds of parametric point processes, *e.g.*, the self- and mutually-correcting processes, we can learn their parameters by gradient descent or stochastic gradient descent (SGD).

3.2. Learning warping/unwarping functions

Given θ^k, we need to update the warping/unwarping functions. To simplify the problem and accelerate our learning method, we take advantage of the warping functions estimated in the previous iteration, *i.e.*, $\{W_m^{k-1}\}_{m=1}^M$, and decompose the problem into M independent problems: for $m = 1, ..., M$, W_m^k is the solution of

$$\begin{aligned} &\min_{W_m} - \log \mathcal{L}(\theta^k; W_m^{-1}(S_m)) \\ &+ \gamma \int_0^T \left| \frac{W_m^{-1}(s)}{M} + \frac{\sum_{m' \neq m} (W_{m'}^{k-1})^{-1}(s)}{M} - s \right|^2 ds \\ &s.t.\ W_m^{-1}(0) = 0,\ W_m^{-1}(T) = T,\ W_m^{-1\prime}(t) > 0. \end{aligned} \quad (5)$$

Solving these problems is non-trivial, requiring further parameterization of the warping functions $\{W_m\}_{m=1}^M$, or equivalently, the unwarping functions $\{W_m^{-1}\}_{m=1}^M$.

We apply a set of piecewise linear models to fit the unwarping functions, for the convenience of mathematical derivation and computation. Specifically, given L landmarks $\{t_1, ..., t_L\}$ in $[0, T]$, where $t_1 = 0$, $t_L = T$ and $t_l < t_{l+1}$, we model W_m^{-1} for $m = 1, ..., M$ as

$$W_m^{-1}(t) = a_l^m t + b_l^m, \text{ if } t \in [t_l, t_{l+1}). \quad (6)$$

Denoting $\boldsymbol{a}^m = \{a_l^m\}_{l=1}^{L-1}$ and $\boldsymbol{b}^m = \{b_l^m\}_{l=1}^{L-1}$ as the parameters of the model, we rewrite the regularizer and the constraints of W_m^{-1} as

$$\begin{aligned} &\int_0^T \left| \frac{W_m^{-1}(s)}{M} + \frac{\sum_{m' \neq m} (W_{m'}^{k-1})^{-1}(s)}{M} - s \right|^2 ds \\ &\to \left\| \frac{1}{M} \boldsymbol{a}^m + \boldsymbol{a}^{\bar{m}} \right\|_2^2 + \left\| \frac{1}{M} \boldsymbol{b}^m + \boldsymbol{b}^{\bar{m}} \right\|_2^2, \\ &W_m^{-1}(0) = 0 \to b_1^m = 0, \\ &W_m^{-1}(T) = T \to a_{L-1}^m T + b_{L-1}^m = T, \\ &W_m^{-1\prime}(t) > 0 \to a_l^m > 0 \text{ for } l = 1, ..., L-1, \end{aligned} \quad (7)$$

where $\| \cdot \|_2$ indicates the ℓ_2 norm of a vector, $\boldsymbol{a}^{\bar{m}} = \frac{\sum_{m' \neq m} \boldsymbol{a}^{m', k-1}}{M} - \mathbf{1}$ and $\boldsymbol{b}^{\bar{m}} = \frac{\sum_{m' \neq m} \boldsymbol{b}^{m', k-1}}{M}$. $\boldsymbol{a}^{m', k-1}$ and $\boldsymbol{b}^{m', k-1}$ are estimated in the previous iteration. To guarantee continuity of W_m^{-1}, we further impose the following constraints on $\boldsymbol{a}^m$ and $\boldsymbol{b}^m$: for $l = 1, ..., L-2$,

$$a_l^m t_{l+1} + b_l^m = a_{l+1}^m t_{l+1} + b_{l+1}^m. \quad (8)$$

Based on the piecewise linear model and the exponential-like intensity assumption, we propose a tight upper bound for the negative log-likelihood in (5):

$$\begin{aligned} &- \log \mathcal{L}(\theta^k; W_m^{-1}(S_m)) \\ =& \sum_{c=1}^C \int_0^T \lambda_c(W_m^{-1}(s)) ds - \sum_{i=1}^{I_m} \log \lambda_{c_i^m}(W_m^{-1}(t_i^m)) \\ \leq& \sum\nolimits_{c=1}^C \int_0^T \lambda_c(s) dW_m(s) \\ &- \sum\nolimits_{i=1}^{I_m} \sum\nolimits_{j=1}^{J_i} q_{ij}^m \log(\lambda_{c_i^m}(W_m^{-1}(t_i^m))/q_{ij}^m) \\ =& \sum_{l=1}^{L-1} \left[\frac{p_l^m}{a_l^m} - \sum_{j=1}^J \sum_{t_i^m \in [t_l, t_{l+1})} q_{ij}^m f_j(a_l^m t_i^m + b_l^m) \right] + \mathsf{C} \\ =& \mathcal{Q}(\boldsymbol{a}^m, \boldsymbol{b}^m). \end{aligned} \quad (9)$$

Here, $\lambda_{c_i^m}(W_m^{-1}(t_i^m)) = \sum_{j=1}^J \exp(f_j(W_m^{-1}(t_i^m); \theta^k))$, the coefficients $p_l^m = \sum_c \int_{W_m^{-1}(t_l)}^{W_m^{-1}(t_{l+1})} \lambda_c(s) ds$, $q_{ij}^m = \frac{\exp(f_j(W_m^{-1}(t_j^m)))}{\lambda_{c_i^m}(W_m^{-1}(t_i^m))}$ and C is the constant independent to W_m^{-1}. The inequality is based on Jensen's inequality and the $\{p_l^m, q_{ij}^m\}$ are calculated based on the parameters estimated in the previous iteration. The detailed derivation and the implementation for Hawkes process are given in Appendix 9.3 and 9.4. Considering (7, 8, 9) together, we propose the surrogate problem of (5):

$$\begin{aligned} &\min_{\boldsymbol{a}^m, \boldsymbol{b}^m} \mathcal{Q}(\boldsymbol{a}^m, \boldsymbol{b}^m) + \gamma \left\| \frac{\boldsymbol{a}^m}{M} + \boldsymbol{a}^{\bar{m}} \right\|_2^2 + \gamma \left\| \frac{\boldsymbol{b}^m}{M} + \boldsymbol{b}^{\bar{m}} \right\|_2^2 \\ &s.t.\ 1)\ b_1^m = 0,\ a_{L-1}^m T + b_{L-1}^m = T, \\ &\quad 2) \text{ for } l = 1, ..., L-1,\ a_l^m > 0, \text{ and} \\ &\quad 3)\ a_l^m t_{l+1} + b_l^m = a_{l+1}^m t_{l+1} + b_{l+1}^m. \end{aligned} \quad (10)$$

(10) is a typical constrained nonlinear programming problem. Many optimization methods can be applied here, $e.g.$, sequential quadratic programming and an interior-point method. Note that after getting optimal $\boldsymbol{a}^m$ and $\boldsymbol{b}^m$, we need to re-calculate the $\{p_l^m, q_{ij}^m\}$ in $\mathcal{Q}$ and solve (10) iteratively until the objective function converges.

Repeating the two steps above, we estimate the model and the warping/unwarping functions effectively.

3.3. Justifiability Analysis

The reasons for applying piecewise linear models to warping functions are twofold. First, our learning method involves the computation of unwarping function W_m^{-1} and the derivative of warping function $W_m^{'}$. Applying our piecewise linear model, both warping and unwarping functions can be represented explicitly. If we use other basis functions, $e.g.$, Gaussian basis, to represent W_m (or W_m^{-1}), the W_m^{-1} (or $W_m^{'}$) may be hard to be represented in closed-form. Second, compared to the finite element analysis used in functional optimization and differential equations, which discretizes functions into many grids, our piecewise linear model requires much fewer parameters, which reduces the risk of over-fitting and the computational complexity.

Complexity Consider a C-dimensional Hawkes process as an example. We implement the MLE step and the updating of unwarping functions via an EM-based framework (Zhou et al., 2013) and an interior-point method (Potra & Wright, 2000), respectively. Given M sequences with I events in each, the computational complexity of our method per iteration in the worst case is $\mathcal{O}(MI^2 + C^2 + ML^3)$. The $\mathcal{O}(MI^2)$ and $\mathcal{O}(C^2)$ correspond to the computational complexity of the E-step and the M-step, and the $\mathcal{O}(ML^3)$ corresponds to the computational complexity of solving M nonlinear programming with $2L$ variables per each in the worst case. Because we update unwarping functions by solving M independent optimization problems in parallel, the time complexity of our method can be $\mathcal{O}(MI^2 + C^2 + L^3)$.

Convergence Our learning method converges in each step. For parametric point processes like Hawkes processes, their likelihood functions are convex and the convergence of the MLE-step is guaranteed. Further, the objective function in (10) is convex, as shown in Appendix 9.5, thus updating of the unwarping functions also converges well.

Compared with existing point process registration methods, $e.g.$, the Wasserstein learning-based registration method (WLR) (Bigot et al., 2012; Panaretos & Zemel, 2016; Zemel & Panaretos, 2017) and the multi-task learning-based method (MTL) (Luo et al., 2015), our RPP method has several advantages. First, both the WLR and the MTL require learning a specific model for each event sequence. For complicated multi-dimensional point processes, they require a large amount of events per sequence to learn reliable models independently, which might be unavailable in practice. Our method has much fewer parameters, and thus has much lower computational complexity and lower risk of overfitting. Second, both the WLR and the MTL decompose the learning of model and warping functions into two independent steps. The estimation error caused in the previous step will propagate to the following one. On the contrary, our method optimizes model and warping functions alternatively with guaranteed convergence, so the estimation error will be suppressed iteratively.

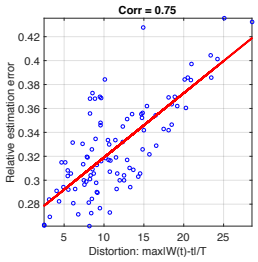


Figure 2. In 100 trials, 40 sequences with length T are generated by a 1D Hawkes process and warped by a function with a certain $\|W(t) - t\|_\infty$. The correlation between the estimation errors and the distortion is 0.75.

4. Potential Improvement Based on Stitching

Empirically, the influence of warped data on learning results is correlated with the distortion of warping function compared to identity function. The distortion should be a measurement not only dependent with the difference between warping function and identity function but also related to the scale of time window because the distortion on a certain scale becomes ignorable when we observe and analyze it on a larger scale. In particular, we propose a definition the distortion as $D = \frac{\|W(t)-t\|_\infty}{T}$ in this work. Here, $\|W(t)-t\|_\infty = \max\{|W(t)-t|, \forall t \in [0,T]\}$, which represents the most serious warping achieved by the warping function, and T is the length of time window. In Fig. 2, we show that the distortion based on this definition is highly correlated with the relative estimation error ($i.e.$, $\frac{\|\theta^*-\theta\|_2}{\|\theta\|_2}$, where θ is the ground truth and θ^* is the estimation result).

This relationship $\frac{\|\theta^*-\theta\|_2}{\|\theta\|_2} \propto D$ implies a potential strategy to further improve our learning results. Suppose that we have two warped sequences $S_1 = \{(t_i^1, c_i^1)\}_{i=1}^{I_1}$ and $S_2 = \{(t_i^2, c_i^2)\}_{i=1}^{I_2}$ observed in the time window $[0,T]$, whose distortions are D_1 and D_2, respectively. If we stitch these two sequences together, $i.e.$, $S = S_1 \cup \S_2 = \{(t_1^1, c_1^1), ..., (t_1^2 + T, c_1^2), ...\}$, the distortion of the new sequence in $[0, 2T]$ will be $D = \frac{1}{2}\max\{D_1, D_2\}$. According to the relationship above, learning from the stitched sequence may help us obtain lower estimation error than learning from the separated two sequences.

Note that for memoryless point processes like Poisson processes, such a stitching-based learning strategy will not

cause model misspecification because the stitched sequence obeys the same model with the original short sequences. However, for more complicated model like Hawkes processes or self-correcting processes, the stitching operation may introduce nonexistent triggering patterns. In such a situation, our stitching-based learning strategy suppresses the influence of warping function while rises the risk of model misspecification as an exchange. Fortunately, as discussed in (Xu et al., 2017a), when the intensity function is exponential-like function, the model misspecification problem is ignorable with a small number of stitching operations. The experiments in the experimental section further verifies the feasibility of our method.

5. Related Work

5.1. Temporal point processes

Point processes have been proven to be useful in many applications, $e.g.$, financial analysis (Bacry et al., 2012), social network analysis (Zhou et al., 2013; Zhao et al., 2015), information system analysis (Xu et al., 2016) and clinical data analysis (Xu et al., 2017b). However, most existing work does not consider registering warped parametric point processes. The methods in (Lewis & Mohler, 2011; Yan et al., 2015) try to estimate time scaling parameters for their point processes, but they are only available for the Hawkes processes whose event sequences share the same linear transformation of time. The work in (Luo et al., 2015) is able to jointly learn different Hawkes processes by multi-task learning, but it does not register its learning results or learn sample-specific warping functions.

5.2. Data registration and model registration

As aforementioned, the idiosyncratic aspects of sequential data may be viewed in terms of a sample-specific "warping" of a common latent phenomena, which can be registered based on learned or predefined transformations. Typical methods include the dynamic time warping (DTW) (Berndt & Clifford, 1994; Moeckel & Murray, 1997) and its variants (Wang et al., 2016; Cuturi & Blondel, 2017; Ramsay & Li, 1998), the self-modeling registration method (SMR) (Gervini & Gasser, 2004), the moment-based method (MBM) (James, 2007), the pairwise curve synchronization method (PACE) (Tang & Müller, 2008), and the functional convex averaging (FCA) method (Liu & Müller, 2004). These methods can be categorized in the same framework – the registered curves and the corresponding warping functions are learned alternatively based on a nonlinear least-squares criterion. Instead of using the Euclidean metric, the recent work in (Srivastava et al., 2011) obtains better data registration results by using the Fisher-Rao metric (FRM). For those nonparametric models like Gaussian processes, warping data is beneficial to improve the robustness of learning methods (Snelson et al., 2004; Cunningham et al., 2012; Snoek et al., 2014; Herlands et al., 2016).

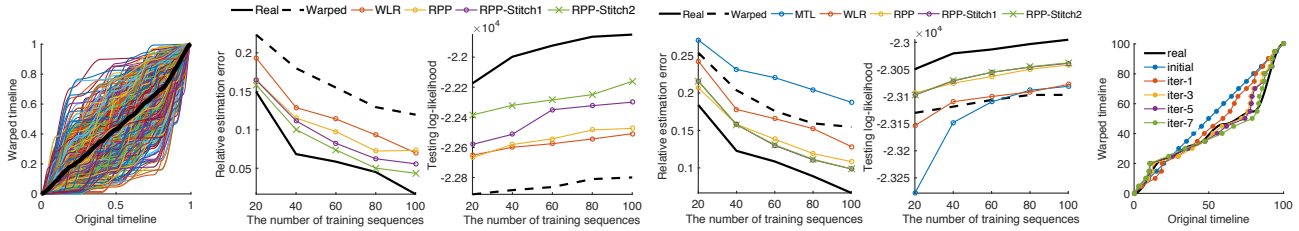
The work in (Panaretos & Zemel, 2016; Zemel & Panaretos, 2017) proposes a model-registration method. Specifically, the unregistered distributions of warped observations are first estimated by nonparametric models, and then the registered point process are estimated as the barycenter of the distributions in Wasserstein space (Muskulus & Verduyn-Lunel, 2011). Finally, the warping function between any unregistered distribution and the registered one is learned as an optimal transport (Anderes et al., 2016). However, all the methods above focus on warping/unwarping continuous curves in a nonparametric way, which are hard to register parametric point processes from idiosyncratic event sequences. The recent combination of Wasserstein learning and neural networks (Arjovsky et al., 2017; Xiao et al., 2017) achieves encouraging improvements on learning robust generative models from imperfect observations. However, the neural network-based model requires many time-consuming simulation steps in the learning phase, and cannot in general learn explicit warping functions.

6. Experiments

Denote our point process registering method and its variant assisted with stitching operation as **RPP** and **RPP-stitch**, respectively To demonstrate the feasibility and effectiveness of the proposed methods, we compare them to existing point process learning and registration methods on both synthetic and real-world datasets. In particular, we compare to the following methods: purely maximum likelihood estimation based on warped observations (**Warped**); the multi-task learning-based method (**MTL**) (Luo et al., 2015); and the Wasserstein learning-based registration method (**WLR**) (Panaretos & Zemel, 2016). Specifically, the MTL method learns specific parametric point processes jointly from warped event sequences with low-rank and sparse regularizers, and averages the learned parameters over all event sequences in the Euclidean space. The WLR is the state-of-the-art model registration method focusing on point processes and their warped event sequences. To apply the WLR method to learn parametric point process models, we first follow the work in (Panaretos & Zemel, 2016), learning the densities of observed events by kernel density estimation (KDE) (Sheather & Jones, 1991), and learning the warping functions by finding the optimal transport between the densities and their barycenter in the Wasserstein space. Finally, we apply the reversed warping functions to unwarp the observations and learn a parametric point process.

6.1. Synthetic data

We simulate an 1D inhomogeneous Poisson process and a 4-dimensional Hawkes process, respectively. For each



(a) $W(t)$ (b) Inhomogeneous Poisson process (c) Hawkes process (d) Convergence

Figure 3. Comparisons for various methods on synthetic data.

synthetic data set, we generate 200 event sequences in the time window $[0, 100]$ using Ogata's thinning method (Ogata, 1981) and divide them equally into a training set and a testing set. The intensity function of the Poisson process is represented as $\sum_{j=1}^{5} \exp_{t_j}(-(t - t_j))$, where t_j is uniformly sampled from $[0, 100]$, while the intensity function of the Hawkes process is defined as the model in (Zhou et al., 2013). Each sequence in the training set is modified by a specific warping function. The warping functions are visualized in Fig. 3(a), in which each color curve represents a warping function and the black bold curve represents the average of all the functions. The generation method of the warping functions is given in Appendix 9.6. It ensures that both the warping and the unwarping functions are monotone increasing and the averaged warping and unwarping functions are close to an identity function.

Given the training data, we can learn registered point process models by different methods and evaluate their performance on 1) the relative estimation error; and 2) the log-likelihood of testing set. For each method, we test it in 5 trials on the two data sets, respectively, and visualize its averaged results in Figs. 3(b) and 3(c). The black bold curves correspond to the MLE based on unwarped data, which achieves the best performance ($i.e.$, the lowest estimation error and the highest log-likelihood), while the black dot curves correspond to the MLE based on warped data. The performance of a good registration method should be much better than the black dot curves and approach to the block bold curves. Our RPP method achieves superior performance to MTL and WLR. The performance of MTL is even worse than that of applying MLE to warped data directly, especially in the case with few training data. This result implies that 1) the sparse and low-rank structure imposed in the multi-task learning phase cannot reflect the actual influence of warped data on the distribution of parameters; and 2) the average of the parameters in the Euclidean space does not converge well to the ground truth. The performance of WLR is comparable to that of applying MLE to warped data directly, which verifies our claim that the WLR is unsuitable for learning complicated point processes when observations are not sufficient enough. Both MTL and WLR reply on a strategy of learning a specific model for each event sequence, and then averaging the models in a predefined space. This strategy ignores a fact that the number of events in a single event sequence is often insufficient to learn a reliable model in practice. Our RPP method, by contrast, learns a single registered model and all warping functions jointly in an iterative manner, rather than in independent steps. As a result, our method suppresses the risk of over-fitting and achieves much better results. Furthermore, we illustrate the learning process of a warping function in Fig. 3(d) and verify the convergence of our RPP method. The black bold curve corresponds to the ground truth and the blue line is the initialization of our estimation. Applying our RPP method, the learning result converges after 7 iterations and the final estimation of the warping function approaches the ground truth.

The usefulness of the stitching strategy is tested as well. In particular, the "RPP-Stitch K" means that for each event sequence we randomly stitch it with K other event sequences and then apply our RPP method to the 200 stitched sequences in the time window $[0, 100(K + 1)]$. We can find that for both Poisson processes and Hawkes processes, "RPP-Stitch 1" obtains better results than original RPP method, which verifies the improvements caused by the stitching strategy. Another advantage of the stitching strategy is improving the stability of our learning results, especially in the cases with small training sets. Given 20 training sequences, the standard deviation of the estimation errors in 5 trial is 0.053 for original RPP, 0.037 for "RPP-Stitch 1" and 0.033 for "RPP-Stitch 2". However, for Poisson processes the improvements can be further enhanced by applying stitching operations multiple times ($i.e.$, $K = 2$), while for Hawkes processes the improvements are almost unchanged. As we discussed in Section 4, applying too many stitching operations to the historically-dependent point processes may cause model misspecification and counteract the benefits from suppressing distortions.

6.2. Real-world data

We test our methods and compare it with the WLR on two real-world datasets: the MIMIC III dataset (Johnson et al., 2016) and the Linkedin dataset (Xu et al., 2017a). The MIMIC III dataset contains over ten thousand patient ad-

mission records over ten years. Each admission record is a sequence, with admission time stamps and the ICD-9 codes of diseases. Following (Xu et al., 2017a), we assume that there are triggering patterns between different diseases, which can be modeled by a Hawkes process. We focus on modeling the triggering patterns between the diseases of the circulatory system, which are grouped into 8 categories. We extract $1,129$ admission records related to the 8 categories as the training set. Each record can be viewed as an event sequence warped from a "standard" record because of the idiosyncratic nature of different patients. For the Linkedin dataset, we extract 709 users having working experience in 7 IT companies. Similarly, the timeline of different users can be different, because they have different working experience and personal conditions, and the status of the job market when they jump is different as well. We want to learn a "standard" Hawkes process to measure the relationships among the companies and exclude these uncertain factors.

We apply different model registration methods to learn registered Hawkes processes from the two real-world datasets. The evaluation is challenging because both the ground truth of the model and that of the warping functions are unknown. Fortunately, we can use learning results to evaluate the risks of under- and over-registration for different methods in an empirical way. Given unwarped event sequences estimated by different methods, we learn the parameter of model θ^* and estimate its variance $var(\theta^*)$ by parametric bootstrapping (Wassermann, 2006). For the method with a lower risk of under-registration, its learning result should be more stable and the estimated variance should be smaller. Therefore, we can use the estimated variance as a metric for the risk of under-registration, $i.e.$, $risk_{under} = var(\theta^*)$. We define the following metric to evaluate the risk of over-registration: $risk_{over} = \frac{\int_0^T |s-\overline{W}(s)|^2 ds}{\frac{1}{M}\sum_{m=1}^M \int_0^T |W_m(s)-\overline{W}(s)|^2 ds}$, where $\overline{W}(s) = \frac{1}{M}\sum_m W_m(s)$. The numerator is the distance between the mean of warping functions and an identity function, and the denominator is the variance of warping functions. When the estimated warping functions have a small variance ($i.e.$, the warping functions are similar to each other) but are very distinct from identity function ($i.e.$, the bias of the warping functions is large), it means that the corresponding method causes over-registration.

The side information of the dataset is also helpful to evaluate the appropriateness of the learning result. In Fig. 4(a), most of the admission records in the MIMIC III dataset are from relatively old patients. The incidence of circulatory system diseases is mainly correlated with patient age. Learning a "standard" patient model from a dataset dominated by old patients, we can imagine that the admission record of an old patient should be more similar to that of the "standard" patient, and the corresponding warping function should be closer to the identity function. Therefore, given the deviations between learned warping functions and the identity function, we can calculate the Kendall's rank correlation between the warping deviations and the ages of the patients. Similarly, in Fig. 4(b), most of samples in the Linkedin dataset are from young users with 4 or fewer working years, so these young users' behaviors should likely be close to that of the "standard" job-hopping model learned from the data, and the warping deviations should be correlated with the working years.

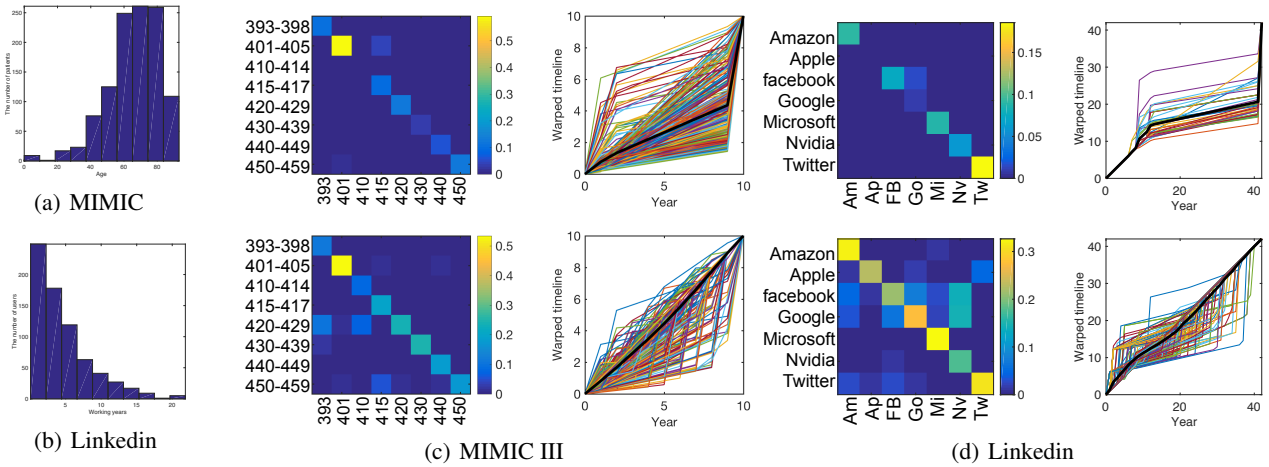
Table 1. Comparison for various methods.

Data	Method	$risk_{under}$	$risk_{over}$	Rank Corr.
MIMIC-III	WLR	0.018	0.055	0.025
	RPP	0.011	0.009	**0.053**
	RPP-Stitch1	**0.003**	**0.002**	**0.053**
LinkedIn	WLR	0.029	0.657	0.344
	RPP	0.025	0.010	0.375
	RPP-Stitch1	**0.005**	**0.006**	**0.387**

Table 1 shows the comparison between our methods (RPP and RPP-Stitch1) and the WLR method on these two datasets. We find that our RPP method outperforms WLR consistently on different metrics and different datasets, obtaining lower risks of under- and over-registration and higher rank correlation. In particular, the low risk of under-registration means that the parameter θ^* learned by our method is stable. The low risk of over-registration means that the warping/unwarping functions we learned have good diversity and low bias. The high rank correlation verifies the justifiability of our method – the warping deviations of dominant samples ($i.e.$, the old patients in MIMIC III and young employees in Linkedin data) are smaller than those of minor samples ($i.e.$, the young patients and the old employees). Similar to the case of synthetic data, applying stitching strategy once, we can further improve the learning results.

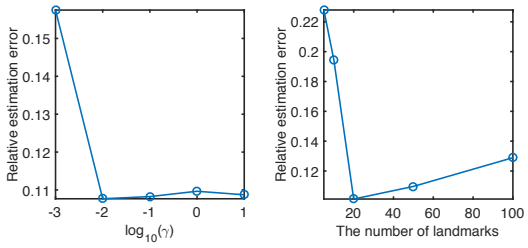
Figures 4(c) and 4(d) compare the infectivity matrices[1] of the registered Hawkes processes and the warping functions learned by WLR and our RPP-Stitch1 for the two datasets. These results further verify the superiority of our method. First, the warping/unwarping functions we learned have good diversity and the bias of the functions is lower than than that of the functions learned by WLR. Second, the infectivity matrices learned by our RPP-Stitch1 are more dense and informative, which reflect some reasonable phenomena that are not found by WLR. For the MIMIC III data, the infectivity matrix of WLR only reflects the self-triggering patterns of the disease categories, while ours is more informative: the 5-th row of our matrix (the bottom-left subfigure in Fig. 4(c)) corresponds to the category "other forms of heart disease" (ICD-9 code 420-429), which contains many

[1]Infectivity matrix is denoted as $\mathbf{\Psi} = [\psi_{cc'}]$. Its element is the integral of impact function over time, $i.e.$, $\psi_{cc'} = \int_0^T \phi_{cc'}(s)ds$.



(a) MIMIC (b) Linkedin (c) MIMIC III (d) Linkedin

Figure 4. Experimental results of our method on real-world datasets. In (c) and (d), the first row corresponds to the infectivity matrix and the warping functions learned by WLR, and the second row corresponds to those learned by our RPP-Stitch1. The black bold curves are the average of warping functions.



(a) γ from 10^{-3} to 10 (b) L from 5 to 100

Figure 5. Illustration of robustness.

miscellaneous heart diseases and has complicated relationships with other categories. Our learning result reflects this fact – the 5-th row of our infectivity matrix contains many non-zero elements. For the Linkedin data, the infectivity matrix of our method reveals more information besides the self-triggering patterns: 1) The values of "Facebook-Google" and "Google-Facebook" imply that job-hopping behaviors happen frequently between Facebook and Google, which reflects fierce competition between these companies. 2) The values of "Facebook-Nvidia" and "Google-Nvidia" reflect the fact that recent years many Nvidia's employees jump to Google and Facebook to develop the hardware of AI. More detailed analysis are given in Appendix 9.7.

6.3. Robustness analysis

We investigate the robustness of our method to variations in its parameters, including the weight of regularizer γ and the number of landmarks L. In particular, we learn models from the synthetic data by our method with different configurations, and visualize the estimation errors with respect to these two parameters in Fig. 5. The weight γ controls the importance of the regularizer, which is correlated with the strictness of the unbiasedness assumption. The larger γ, the more similarity we have between unwarping function and identity function. In Fig. 5(a) we find that our method is robust to the change of γ in a wide range ($i.e.$, from 10^{-3} to 1). When γ is too small ($i.e.$, $\gamma = 10^{-3}$), however, the estimation error increases because the regularizer is too weak to prevent over-registration. The number of landmarks L has an effect on the representation power of our method. In Fig. 5(b), we find that the lowest estimation error is achieved when the number of landmarks $L = 20$. When L is too small, our piecewise linear model is over-simplified and cannot fit complicated warping functions well. When L is too large, (10) has too many variables and the updating of warping function suffers to the problem of over-fitting.

7. Conclusions and Future work

We have proposed an alternating optimization method to learn parametric point processes from idiosyncratic observations. We demonstrate its justifiability and advantages relative to existing methods. Additionally, we also consider the influence of the stitching operation on the learning results and show the potential benefits empirically. Our method has potentials to many applications, including admission data analysis and job-hopping behavior analysis. In the future, we plan to extend our method to more complicated point process models and analyze the influence of the stitching operation theoretically.

8. Acknowledgment

This work was supported in part by DARPA, DOE, NIH, ONR, NSF, IIS-1717916 and CMMI-1745382. We thank Yoav Zemel for discussions and comments about this work.

References

Anderes, E., Borgwardt, S., and Miller, J. Discrete Wasserstein barycenters: optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2): 389–409, 2016.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Bacry, E., Dayri, K., and Muzy, J.-F. Non-parametric kernel estimation for symmetric Hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):1–12, 2012.

Berndt, D. J. and Clifford, J. Using dynamic time warping to find patterns in time series. In *KDD workshop*, 1994.

Bigot, J., Klein, T., et al. Consistent estimation of a population barycenter in the Wasserstein space. *ArXiv e-prints*, 2012.

Cunningham, J., Ghahramani, Z., and Rasmussen, C. E. Gaussian processes for time-marked time-series data. In *AISTATS*, 2012.

Cuturi, M. and Blondel, M. Soft-dtw: a differentiable loss function for time-series. *arXiv preprint arXiv:1703.01541*, 2017.

Daley, D. J. and Vere-Jones, D. *An introduction to the theory of point processes: volume II: general theory and structure*, volume 2. Springer Science & Business Media, 2007.

Gervini, D. and Gasser, T. Self-modelling warping functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):959–971, 2004.

Hawkes, A. G. and Oakes, D. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Herlands, W., Wilson, A., Nickisch, H., Flaxman, S., Neill, D., Van Panhuis, W., and Xing, E. Scalable gaussian processes for characterizing multidimensional change surfaces. In *AISTATS*, 2016.

Isham, V. and Westcott, M. A self-correcting point process. *Stochastic Processes and Their Applications*, 8(3):335–347, 1979.

James, G. M. Curve alignment by moments. *The Annals of Applied Statistics*, pp. 480–501, 2007.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016.

Lewis, E. and Mohler, G. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011.

Liu, X. and Müller, H.-G. Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99(467):687–699, 2004.

Luo, D., Xu, H., Zhen, Y., Ning, X., Zha, H., Yang, X., and Zhang, W. Multi-task multi-dimensional Hawkes processes for modeling event sequences. In *IJCAI*, 2015.

Mammen, E. Nonparametric estimation of locally stationary hawkes processe. *arXiv preprint arXiv:1707.04469*, 2017.

Moeckel, R. and Murray, B. Measuring the distance between time series. *Physica D: Nonlinear Phenomena*, 102(3-4): 187–194, 1997.

Muskulus, M. and Verduyn-Lunel, S. Wasserstein distances in the analysis of time series and dynamical systems. *Physica D: Nonlinear Phenomena*, 240(1):45–58, 2011.

Ogata, Y. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

Panaretos, V. M. and Zemel, Y. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2): 771–812, 2016.

Potra, F. A. and Wright, S. J. Interior-point methods. *Journal of Computational and Applied Mathematics*, 124(1): 281–302, 2000.

Ramsay, J. O. and Li, X. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363, 1998.

Roueff, F., Von Sachs, R., and Sansonnet, L. Locally stationary hawkes processes. *Stochastic Processes and their Applications*, 126(6):1710–1743, 2016.

Sheather, S. J. and Jones, M. C. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690, 1991.

Snelson, E., Ghahramani, Z., and Rasmussen, C. E. Warped gaussian processes. In *NIPS*, 2004.

Snoek, J., Swersky, K., Zemel, R., and Adams, R. Input warping for bayesian optimization of non-stationary functions. In *ICML*, 2014.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. Registration of functional data using Fisher-Rao metric. *arXiv preprint arXiv:1103.3817*, 2011.

Tang, R. and Müller, H.-G. Pairwise curve synchronization for functional data. *Biometrika*, 95(4):875–889, 2008.

Wang, Y., Miller, D. J., Poskanzer, K., Wang, Y., Tian, L., and Yu, G. Graphical time warping for joint alignment of multiple curves. In *NIPS*, 2016.

Wassermann, L. *All of nonparametric statistics*. Springer Science+ Business Media, New York, 2006.

Xiao, S., Farajtabar, M., Ye, X., Yan, J., Song, L., and Zha, H. Wasserstein learning of deep generative point process models. *arXiv preprint arXiv:1705.08051*, 2017.

Xu, H., Zhen, Y., and Zha, H. Trailer generation via a point process-based visual attractiveness model. In *IJCAI*, 2015.

Xu, H., Farajtabar, M., and Zha, H. Learning Granger causality for Hawkes processes. In *ICML*, 2016.

Xu, H., Luo, D., and Zha, H. Learning Hawkes processes from short doubly-censored event sequences. In *ICML*, 2017a.

Xu, H., Wu, W., Nemati, S., and Zha, H. Patient flow prediction via discriminative learning of mutually-correcting processes. *IEEE transactions on Knowledge and Data Engineering*, 29(1):157–171, 2017b.

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., Chu, S., and Yang, X. On machine learning towards predictive sales pipeline analytics. In *AAAI*, 2015.

Zemel, Y. and Panaretos, V. M. Fréchet means and Procrustes analysis in Wasserstein space. *arXiv preprint arXiv:1701.06876*, 2017.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *KDD*, 2015.

Zhou, K., Zha, H., and Song, L. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *AISTATS*, 2013.

9. Appendix

9.1. Exponential-like intensity functions

We given some typical and important point processes with exponential-like intensity functions, *i.e.*, $\lambda(t) = \sum_j \exp_{t_j}(f(t;\theta,\mathcal{H}_t^C))$. More specifically, for Hawkes processes and self-correcting processes, this formulation can be further rewritten as $\lambda(t) = \sum_j \alpha_j \exp(\beta_j t)$. For the convenience of expression, we only consider 1-D point processes, *i.e.*, the number of event types $C = 1$, but these examples can be easily extended to multi-dimensional cases.

Hawkes processes. The intensity function of a 1-D Hawkes process is

$$\lambda(t) = \mu + \sum\nolimits_{t_i<t} \phi(t - t_i), \tag{11}$$

A typical implementation of the impact function $\phi(t)$ is exponential function, *i.e.*, $\rho \exp(-wt)$ in (Hawkes & Oakes, 1974; Lewis & Mohler, 2011; Zhou et al., 2013; Yan et al., 2015). Therefore, we can rewrite (11) as

$$\begin{aligned}\lambda(t) &= \mu + \sum\nolimits_{t_i<t} \phi(t - t_i) \\ &= \mu \exp(0t) + \sum\nolimits_{t_i<t} \rho \exp(wt_i)\exp(-wt)) \\ &= \sum\nolimits_{j=1}^{J} \alpha_j \exp(-\beta_j t),\end{aligned} \tag{12}$$

where $J = 1 + |\{t_i : t_i < t\}|$. We can find that for $j = 1$, $\beta_j = 0$ and $\alpha_j = \mu$; for $j = 2, .., J$, $\beta_j = w$ and $\alpha_j = \rho \exp(wt_i)$.

Self-correcting processes. The intensity function of a 1-D self-correcting process (Isham & Westcott, 1979; Xu et al., 2015) is

$$\lambda(t) = \exp(\mu t - \sum\nolimits_{t_i<t} \phi(t_i)). \tag{13}$$

Generally, $\phi(t)$ can be 1) a linear function of time, *i.e.*, $\phi(t) = \rho t$; or 2) a constant, *i.e.*, $\phi(t) = \rho$. In this case, we can simply represent $\lambda(t)$ as an exponential function $\alpha \exp(-\beta t)$, where $\alpha = \exp(-\sum_{t_i<t} \phi(t_i))$ and $\beta = -\mu$.

9.2. The proof of Theorem 2.1

Proof. Denote an original (unwarped) event sequence as D. The negative log-likelihood function of the target point process N_θ can be written as

$$-\log \mathcal{L}(\theta; D) = \int_0^T \lambda(s)ds - \sum_i \log \lambda(t_i), \tag{14}$$

where t_i is the i-th event of the sequence D. When the training sequence D is warped by a warping function $W : [0, T] \mapsto [0, T]$ and the warping function is continuous and differentiable (almost everywhere), we have

$$\begin{aligned}&-\log \mathcal{L}(\theta; S) \\ &= \int_0^T \lambda(W(s))ds - \sum_i \log \lambda(W(t_i)) \\ &= \int_0^T \lambda(s)dW^{-1}(s) - \sum_i \log \lambda(W(t_i)),\end{aligned} \tag{15}$$

where S is the warped data.

Sufficiency. When the target point process is a homogeneous Poisson process, i.e., $\lambda(t) = \mu$, we can find that

$$-\log \mathcal{L}(\theta; S) = -\log \mathcal{L}(\theta; D) = T\mu - I \log \mu, \tag{16}$$

where I is the number of events. Therefore, both $\hat{\theta}^*$ and $\hat{\theta}$ are equal to $\frac{I}{T}$.

When we relax the range of $W(t)$ but assume that it is a translation, i.e., $W(t) = t + \tau$, the relative distance between arbitrary two events, i.e., $t_i - t_j = W(t_i) - W(t_j)$, is unchanged. Based on the stationarity of the target point process, the learning result is unchanged as well.

Necessity. When the target point process has exponential-like intensity function, the negative log-likelihood is a convex function of θ. The warping function does not change the convexity of the negative log-likelihood. Therefore, when $\hat{\theta}^* = \hat{\theta}$, we have

$$\frac{\partial - \log \mathcal{L}(\theta; S)}{\partial \theta}\Big|_{\hat{\theta}^*} = 0, \tag{17}$$

for the target point process.

Even in the simplest case, i.e., the intensity is a single exponential function $\lambda(t) = \alpha_\theta \exp(-\beta t)$ and only α_θ is a single coefficient related to the parameter θ, we have

$$\begin{aligned}&-\log \mathcal{L}(\theta; S) \\ &= -\log \mathcal{L}(\theta; D) + \int_0^T (1 - (W^{-1})'(s))\lambda(s)ds \\ &\quad - \sum_i \log \frac{\lambda(W(t_i))}{\lambda(t_i)} \\ &= -\log \mathcal{L}(\theta; D) + \alpha_\theta \int_0^T (1 - (W^{-1})'(s)) \exp(-\beta s)ds \\ &\quad - \sum_i \log \frac{\exp(-\beta W(t_i))}{\exp(-\beta t_i)}.\end{aligned}$$

Here, we have

$$\frac{\partial - \log \mathcal{L}(\theta; D)}{\partial \theta}\Big|_{\hat{\theta}^*} = 0, \tag{18}$$

and the last term $-\sum_i \log \frac{\exp(-\beta W(t_i))}{\exp(-\beta t_i)}$ is a constant with respect to θ, therefore, $\frac{\partial - \log \mathcal{L}(\theta; S)}{\partial \theta}|_{\hat{\theta}^*} = 0$ is equivalent

to $\int_0^T (1-(W^{-1})'(s))\exp(-\beta s)ds \equiv 0$ for all kinds of event sequences. This condition satisfies in two situations: 1) $(W^{-1})'(s) \equiv 1$, which corresponds to a translation function; 2) $\beta = 0$, such that $\lambda(t) = \alpha_\theta$ is a constant, which corresponds to a homogeneous Poisson process. □

9.3. The derivation of (9)

Based on the assumption 3 of the target point process, the negative log-likelihood in (5) can be rewrite as

$$\begin{aligned}
&-\log \mathcal{L}(\theta^k; W_m^{-1}(S_m)) \\
=&\sum_{c=1}^{C}\int_0^T \lambda_c(W_m^{-1}(s))ds - \sum_{i=1}^{I_m}\log\lambda_{c_i^m}(W_m^{-1}(t_i^m)) \\
=&\sum\nolimits_{c=1}^{C}\int_{W_m^{-1}(0)}^{W_m^{-1}(T)} \lambda_c(s)dW_m(s) \\
&-\sum\nolimits_{i=1}^{I_m}\log\Bigg(\sum\nolimits_{j=1}^{J_i}\alpha_j\exp(-\beta_j W_m^{-1}(t_i^m))\Bigg) \qquad (19)\\
=&\sum\nolimits_{c=1}^{C}\int_0^T \lambda_c(s)dW_m(s) \\
&-\sum\nolimits_{i=1}^{I_m}\log\Bigg(\sum\nolimits_{j=1}^{J_i}\alpha_j\exp(-\beta_j W_m^{-1}(t_i^m))\Bigg) \\
=&\mathcal{A}+\mathcal{B}.
\end{aligned}$$

On one hand, based on the piecewise linear model of W_m^{-1}, the term $\mathcal{A}$ can be further rewritten as

$$\begin{aligned}
\mathcal{A} =&\sum\nolimits_{c=1}^{C}\int_{W_m^{-1}(0)}^{W_m^{-1}(T)}\lambda_c(s)dW_m(s) \\
=&\sum\nolimits_{c=1}^{C}\sum\nolimits_{l=1}^{L-1}\int_{W_m^{-1}(t_l)}^{W_m^{-1}(t_{l+1})}\lambda_c(s)\frac{dW_m(s)}{ds}ds \qquad (20)\\
=&\sum\nolimits_{l=1}^{L-1}\underbrace{\frac{1}{a_l^m}}_{W_m'}\underbrace{\sum\nolimits_{c=1}^{C}\int_{W_m^{-1}(t_l)}^{W_m^{-1}(t_{l+1})}\lambda_c(s)ds}_{p_l^m}.
\end{aligned}$$

On the other hand, given current estimated parameters, we can calculate

$$\begin{aligned}
q_{ij}^m &= \frac{\alpha_j\exp(-\beta_j W_m^{-1}(t_j^m))}{\sum_{j'}\alpha_{j'=1}^{J_i}\exp(-\beta_{j'}W_m^{-1}(t_{j'}^m))} \qquad (21)\\
&= \frac{\alpha_j\exp(-\beta_j W_m^{-1}(t_j^m))}{\lambda_{c_i^m}(W_m^{-1}(t_i^m))},
\end{aligned}$$

and then apply Jensen's inequality to the term $\mathcal{B}$:

$$\begin{aligned}
\mathcal{B} =& -\sum\nolimits_{i=1}^{I_m}\log\Bigg(\sum\nolimits_{j=1}^{J_i}\alpha_j\exp(-\beta_j W_m^{-1}(t_i^m))\Bigg) \\
\le&\sum\nolimits_{i=1}^{I_m}\sum\nolimits_{j=1}^{J_i} q_{ij}^m\log\frac{q_{ij}^m}{\alpha_j\exp(-\beta_j W_m^{-1}(t_i^m))} \\
=&\sum\nolimits_{i=1}^{I_m}\sum\nolimits_{j=1}^{J_i} q_{ij}^m\Big(\log\frac{q_{ij}^m}{\alpha_j}+\beta_j W_m^{-1}(t_i^m)\Big) \qquad (22)\\
=&\sum_{l=1}^{L-1}\sum_{t_i^m\in[t_l,t_{l+1})}\sum_{j=1}^{J_i} q_{ij}^m\beta_j(a_l^m t_i^m + b_l^m))) + \mathsf{C}
\end{aligned}$$

9.4. Practical implementations

Taking a **multi-dimensional Hawkes process** as an example, we give the implementation details of our learning method. Specifically, the intensity function of the type-c event at time t is

$$\lambda_c(t) = \mu_c + \sum_{t_i<t}\phi_{c_i c_j}\exp(-w(t-t_i)), \qquad (23)$$

where the parameter set θ consists of the background intensity vector $\boldsymbol{\mu} = [\mu_c]$ and the infectivity matrix $\boldsymbol{\Phi} = [\phi_{cc'}]$.

Maximum likelihood. Given unwarped sequences $\{W_m^{-1}(S_m)\}_{m=1}^M$, we can maximize the likelihood of the sequences by an EM-based method (Lewis & Mohler, 2011; Zhou et al., 2013). Specifically, the negative likelihood function and its tight upper bound can be written as

$$\begin{aligned}
&-\sum_{m=1}^{M}\log\mathcal{L}(\theta; W_m^{-1}(S_m)) \\
=&\sum_{m=1}^{M}\Big[\sum_{c=1}^{C}\int_0^T\lambda_c(W_m^{-1}(s))ds \\
&-\sum_{i=1}^{I_m}\log\lambda_{c_i^m}(W_m^{-1}(t_i^m))\Big] \\
=&\sum_{m=1}^{M}\Big[\sum_{c=1}^{C}\Big(T\mu_c+\sum_{i=1}^{I_m}\phi_{cc_i^m}\int_0^{T-t_i^m}\exp(-wW_m^{-1}(s))ds\Big) \\
&-\sum_{i=1}^{I_m}\log\Big(\mu_{c_i^m}+\sum_{j=1}^{i-1}\phi_{c_i^m c_j^m}\exp(-w\tau_{ij})\Big)\Big] \\
\le&\sum_{m=1}^{M}\Big[\sum_{c=1}^{C}\Big(T\mu_c+\sum_{i=1}^{I_m}\phi_{cc_i^m}\int_0^{T-t_i^m}\exp(-wW_m^{-1}(s))ds\Big) \\
&-\sum_{i=1}^{I_m}\Big(p_i\log\frac{\mu_{c_i^m}}{p_i}+\sum_{j=1}^{i-1}p_{ij}\log\frac{\phi_{c_i^m c_j^m}\exp(-w\tau_{ij})}{p_{ij}}\Big)\Big] \\
=&\mathcal{L}(\theta|\hat{\theta}).
\end{aligned}$$

Here, $\tau_{ij} = W_m^{-1}(t_i^m) - W_m^{-1}(t_j^m)$ and $\hat{\theta}$ is current esti-

mated parameters used to calculate $\{p_i, p_{ij}\}$ as

$$\begin{aligned} p_i &= \frac{\hat{\mu}}{\hat{\lambda}_{c_i^m}(W_m^{-1}(t_i^m))}, \\ p_{ij} &= \frac{\hat{\phi}_{c_i^m c_j^m} \exp(-w\tau_{ij})}{\hat{\lambda}_{c_i^m}(W_m^{-1}(t_i^m))}. \end{aligned} \tag{24}$$

As a result, we can update θ by minimizing $\mathcal{L}(\theta|\hat{\theta})$, which has the following closed-form solution:

$$\begin{aligned} \mu_c &= \frac{\sum_m \sum_{c_i^m=c} p_i}{MT}, \\ \phi_{cc'} &= \frac{\sum_m \sum_{c_i^m=c} \sum_{c_j^m=c'} p_{ij}}{\sum_m \sum_{c_i^m=c'} \int_0^{T-t_i^m} \exp(-wW_m^{-1}(s))ds}. \end{aligned} \tag{25}$$

According to the updated parameters, we can go back to calculate $\{p_i, p_{ij}\}$. Repeating the steps above till the objective function ($i.e.$, the negative log-likelihood) converges, we can obtain the optimum model given current $\{W_m\}_{m=1}^M$.

Learning unwarping functions. The key of this step is calculating the $\{p_l^m, q_{ij}^m\}$ mentioned in (20, 22). For p_l^m, we have

$$\begin{aligned} p_l^m =& \sum_{c=1}^{C} \int_{W_m^{-1}(t_l)}^{W_m^{-1}(t_{l+1})} \lambda_c(s)ds \\ =& \sum_{\substack{c=1,\dots,C \\ t_i^m \in [t_l, t_{l+1})}} \Big(\phi_{cc_i^m} \int_0^{W_m^{-1}(t_{l+1}) - W_m^{-1}(t_i^m)} e^{-ws} ds \\ & + \mu_c (W_m^{-1}(t_{l+1}) - W_m^{-1}(t_l)) \Big) \\ =& \sum_{\substack{c=1,\dots,C \\ t_i^m \in [t_l, t_{l+1})}} \Big(\phi_{cc_i^m} \frac{1 - e^{-wa_l^m(t_{l+1} - t_i^m)}}{w} \\ & + \mu_c a_l^m (t_{l+1} - t_l) \Big). \end{aligned} \tag{26}$$

For q_{ij}^m, because

$$\begin{aligned} & \lambda_{c_i^m}(W_m^{-1}(t_i^m)) \\ =& \mu_{c_i^m} + \sum_{j=1}^{i-1} \phi_{c_i^m c_j^m} \exp(-w(W_m^{-1}(t_i^m) - W_m^{-1}(t_j^m))) \\ =& \sum_{j=0}^{i-1} \alpha_j \exp(-\beta_j W_m^{-1}(t_i^m)), \end{aligned} \tag{27}$$

where for $j = 0$, $\alpha_j = \mu_{c_i^m}$ and $\beta_j = 0$; and for $j > 0$, $\alpha_j = \phi_{c_i^m c_j^m} \exp(wW_m^{-1}(t_j^m))$ and $\beta_j = w$, we have

$$q_{ij}^m = \frac{\alpha_j \exp(-\beta_j W_m^{-1}(t_i^m))}{\lambda_{c_i^m}(W_m^{-1}(t_i^m))} \text{ for } j = 0, ..., i-1. \tag{28}$$

In our experiments, we configure our learning algorithm as follows. The number of landmarks $L = 20$. The weight of regularizer $\gamma = 0.01$. The maximum number of outer iteration is 7. The maximum number of inner iteration for learning the Hawkes process model is 15. The maximum number of inner iteration for updating warping functions is 5. The interior-point method is applied.

9.5. The convexity of (10)

Ignoring constraints, (10) can be decomposed into $2(L-1)$ problems with respect to each a_l^m and b_l^m. The objective function in (10) that is related to a_l^m can be formulated as

$$f(x) = \frac{\alpha}{x} + \beta x + (x + \tau)^2, \tag{29}$$

where the unknown variable $x > 0$, the coefficients α and β are nonnegative, and τ is arbitrary. Because when $x > 0$, $\frac{\alpha}{x}$, βx and $(x+\tau)^2$ are convex functions, their sum, $i.e.$, $f(x)$, is also a convex function as well. Similarly, the objective function in (10) that is related to b_l^m can be formulated as

$$f(x) = \beta x + (x + \tau)^2, \tag{30}$$

which is also a convex function.

9.6. Generating warping/unwarping functions

For the synthetic data used in our experiments, each warping function in $[0, T]$ is represented by a set of local cosine basis as

$$\begin{aligned} W_m(t) &= \sum_{n=1}^{N} w_n^m f_n(t), \\ f_n(t) &= \begin{cases} \cos^2(\frac{\pi}{2\Delta}(t - t_n)), & |t - t_n| <= \Delta \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{31}$$

The time window $[0, T]$ is segmented by N landmarks $\{t_n\}_{n=1}^N$, where $t_1 = 0$ and $t_N = T$. For each $f_n(t)$, the landmark t_n is its center and Δ is the distance between adjacent landmarks. The first $N-1$ coefficients $\{w_n^m\}_{n=1}^{N-1}$ is sampled from $[0, T]$ uniformly and sorted by ascending order. The last coefficient w_N^m is set to be T. Using this method, we can ensure that all warping functions (and the corresponding unwarping functions) are monotone increasing maps from $[0, T]$ to $[0, T]$ and their average is close to an identity function.

9.7. Details of experiments

For the MIMIC data set, each admission is associated with a set of diagnose. Based on the priority assigned to the diagnose, we only keep the ICD-9 code with the highest priority as the event type of the admission. In our work, we assume that the admission behaviors of all patients happen from 2001 to 2012 or their death date. In this case, the length

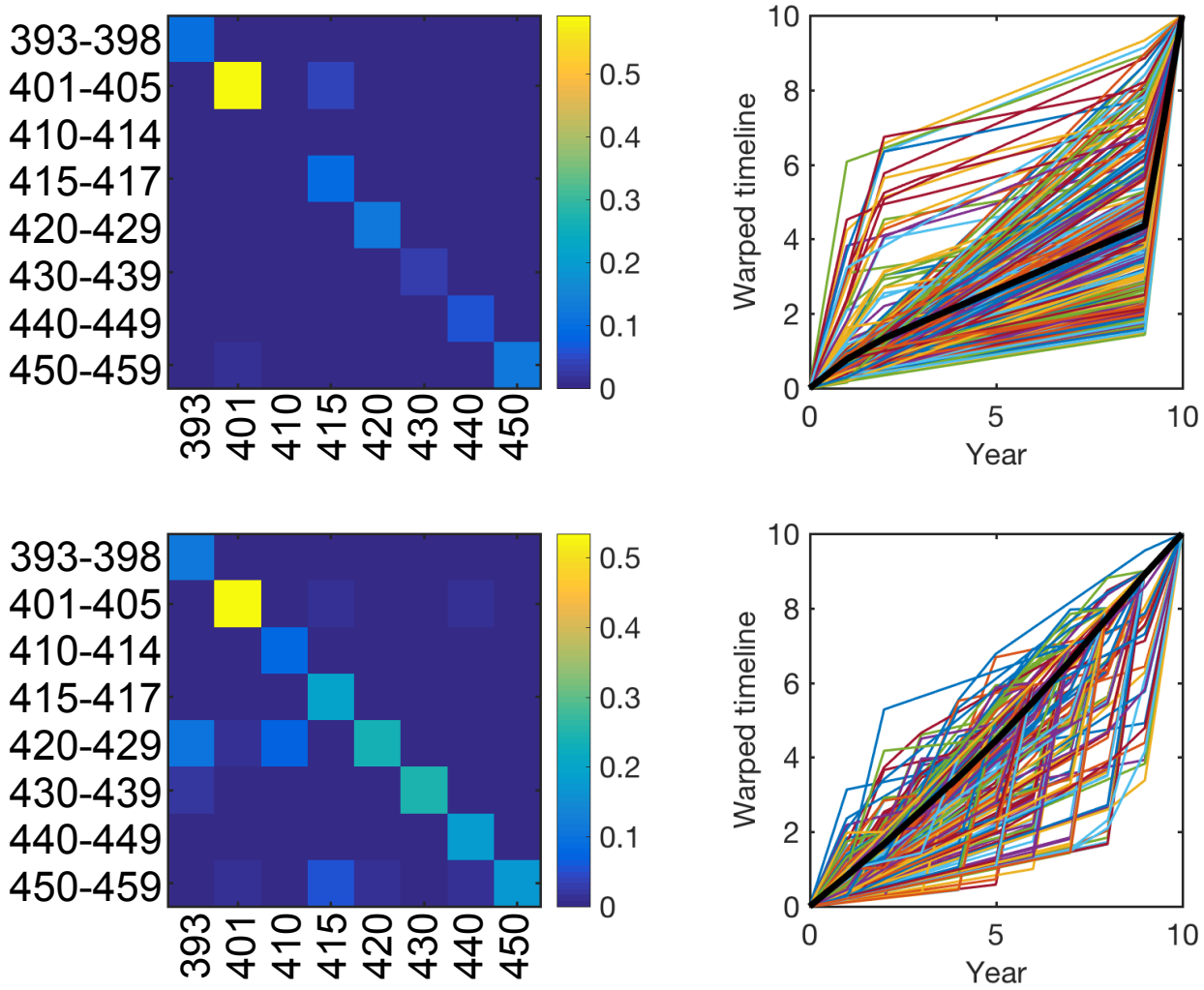


Figure 6. Experimental result of WLR (top) and our method (bottom) on MIMIC III data.

of time window T is different for each patient. Our learning method can be extended to adjust this situation. In particular, we can use specific T's for different event sequences, $i.e.$, replacing T to T^m in our model and learning algorithm. For our piecewise linear model, the distance between adjacent landmarks can be adjusted as well according to T^m. For each patient in the MIMIC data set, we can set the time stamp of its last admission event as T^m.

The categories of the diseases of circulatory system are shown below:

1. Chronic rheumatic heart disease (ICD-9: 393 - 398)
2. Hypertensive disease (ICD-9: 401 - 405)
3. Ischemic heart disease (ICD-9: 410 - 414)
4. Diseases of pulmonary circulation (ICD-9: 415 - 417)
5. Other forms of heart disease (ICD-9: 420 - 429)
6. Cerebrovascular disease (ICD-9: 430 - 438)
7. Diseases of arteries, arterioles, and capillaries (ICD-9: 440 - 449)
8. Diseases of veins and lymphatics, and other diseases of circulatory system (451 - 459)

Using our RPP method, we learn a 8-dimensional Hawkes process from $1,129$ patient's admission records. Compared to synthetic data, the MIMIC III dataset is sparse ($i.e.$, most of the patients have just 2 - 5 admission events), so we use a larger weight for regularizer ($i.e.$, $\gamma = 10$) and fewer landmarks ($i.e.$, $L = 5$). Similarly, we can learn a 7-dimensional Hawkes process from 709 users' job hopping records, in which we also set $\gamma = 10$ and $L = 5$.

These infectivity matrices further verify the justifiability of our learning method because they reflect some reasonable phenomena. In Fig. 6, we can find that:

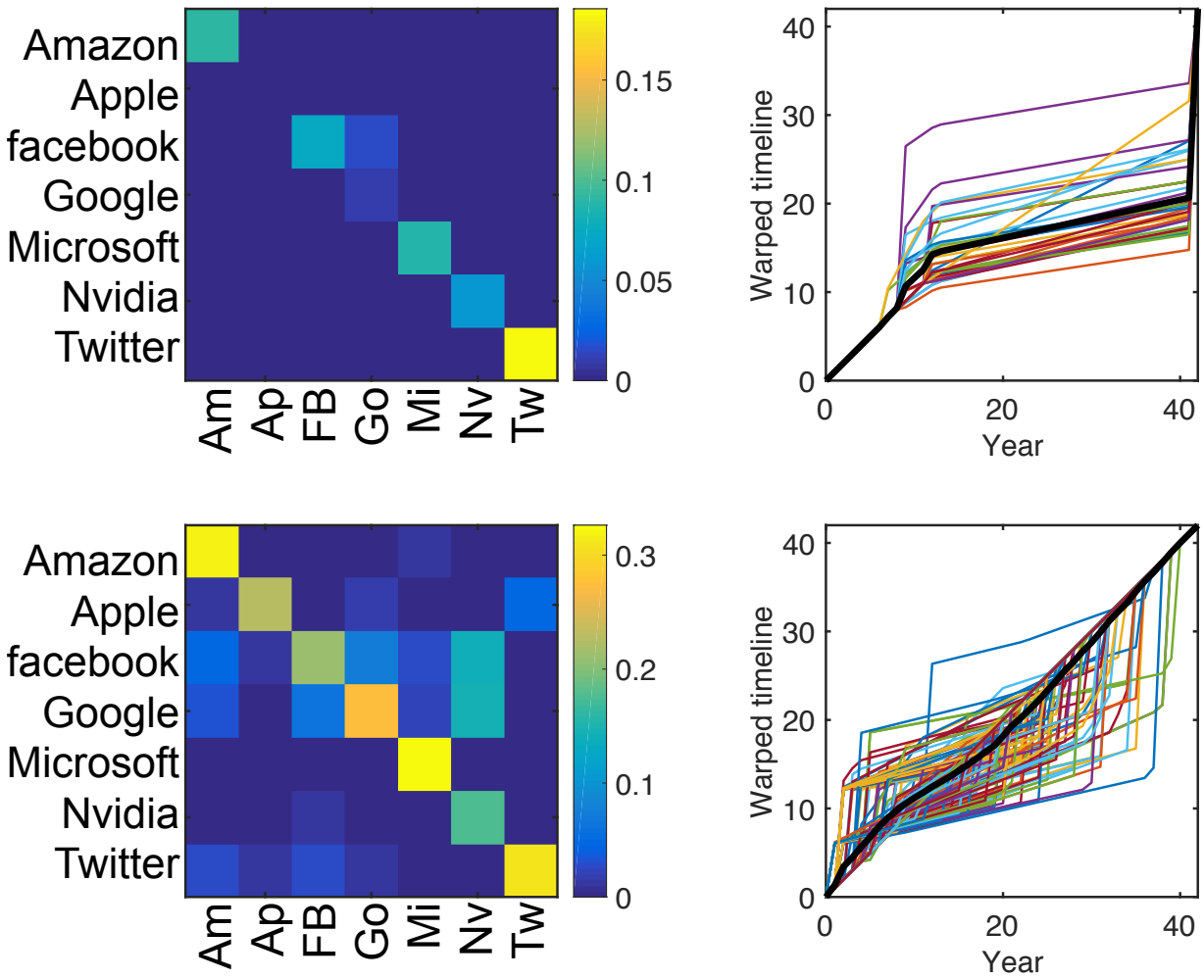


Figure 7. Experimental result of WLR (top) and our method (bottom) on LinkedIn data.

1. All disease categories have strong self-triggering patterns. The "hypertension disease" (ICD-9 code 404-405), which is one of the most common disease in modern society, has the strongest self-triggering pattern — the value of the second diagonal element is over 0.5 It means that for a patient suffering to a certain disease of circulatory system, he or she is likely to re-admit to hospital in next 10 years for the same disease.

2. The 5-th row in Fig. 6 corresponds to the category "other forms of heart disease" (ICD-9 code 420-429). According to its name we can know that this category contains many miscellaneous heart diseases and should have complicated relationships with other categories. Our learning result reflects this fact — the 5-th row of our infectivity matrix contains many non-zero elements, which means that this disease category can be triggered by other disease categories.

In Fig. 7, we can find that:

1. All IT companies have strong self-triggering patterns, which means that most of employees are satisfied to their companies. Especially for Amazon and Microsoft, their diagonal elements are over 0.3. It means that the expected happening rate of internal promotion for their employees is about 0.3 event per year.

2. The elements of "Facebook-Google" and "Google-Facebook" pairs are with high values, which means that job hopping happens frequently between Facebook and Google. This result reflects their fierce competition.

3. The elements of "Facebook-Nvidia" and "Google-Nvidia" are with high values, which reflects the fact that recent years many Nvidia's employees jump to Google and Facebook to develop hardware and systems of AI.

In our opinion, there are three reasons for the increased performance. Firstly, our parametric model is more robust to data insufficiency, which can capture complicated mechanism of event sequences from relatively fewer observations. Secondly, we learn the registered model and the warping functions in an alterative, rather than independent way, to avoid serious model misspecification, and such a method has a good convergence. Thirdly, the proposed piecewise linear model has a good capability to describe warping function approximately, which achieves a trade-off between the complexity of the model and the performance.