A Multitask Point Process Predictive Model

Wenzhao Lian\textsuperscript{1}, Ricardo Henao\textsuperscript{1}, Vinayak Rao\textsuperscript{2}, Joseph Lucas\textsuperscript{3}, Lawrence Carin\textsuperscript{1}

\textsuperscript{1}: Department of Electrical and Computer Engineering, Duke University
\textsuperscript{2}: Department of Statistics, Purdue University
\textsuperscript{3}: Center for Predictive Medicine, Duke Clinical Research Institute

July 8, 2015
Motivation

Point process observations: $\{y_n^u\}, u = 1, \cdots, U$ and $n = 1, \cdots, D^u$

$y_n^u$: the time stamp of the $n$-th arrival of subject/task $u$.

Running example

- Electronical health records
- $\{y_n^u\}_{n=1}^{D^u}$ is one hospital-visit trajectory for a single patient $u$

Figure: Problem demonstration
Motivation

Point process observations: \( \{y^u_n\}, \ u = 1, \ldots, U \) and \( n = 1, \ldots, D^u \)

\( y^u_n \): the time stamp of the \( n \)-th arrival of subject/task \( u \).

Example questions:

- when will the next event happen?
- how many events will happen in \([t, t + L]\)?
Motivation

Point process observations: \( \{y_n^u\}, \ u = 1, \cdots, U \) and \( n = 1, \cdots, D^u \)

\( y_n^u \): the time stamp of the \( n \)-th arrival of subject/task \( u \).

Practical applications:

- Purchasing behavior for individual customers
- Failures in distributed computer systems

Figure: Problem demonstration
Motivation

Task: model the trajectories and make **predictions** for future event arrivals.

Key issues:

- If independently modeling $\rightarrow$ scarce training data per stream
- If sharing across population $\rightarrow$ not aligned in time

![Diagram of Problem Demonstration](image)

**Figure**: Problem demonstration
Agenda

1. Motivation
2. Model
3. Inference
4. Experiments
Conditional Intensity Models

Observations: \( \{y_n^u\}, \ u = 1, \cdots, U \) and \( n = 1, \cdots, D^u \)

\( y_n^u \): the time stamp of the \( n \)-th arrival of subject/task \( u \).

The likelihood in conditional intensity models is defined as [Gunawardana et al., 2011],

\[
p(Y^u) = \prod_{n=1}^{D^u} \gamma^u(y_n^u | h^u(y_n^u)) \exp(- \int_{y_n^u}^{y_{n+1}^u} \gamma^u(\tau | h^u(\tau)) d\tau).
\]

\( h(t) \) summarizes the history of past observations;

\( \gamma(t|h(t)) \): intensity function;

\( N(t + \Delta) - N(t) \sim Ber(\Delta \gamma(t|h(t))) \), as \( \Delta \to 0 \).
Assuming piecewise-constant feature functions $h(t)$:

$$
\begin{align*}
\Delta_i^u &= t_{i+1}^u - t_i^u \\
N^u &: \text{the number of change points of the feature function } h^u(t).
\end{align*}
$$
Assuming piecewise-constant feature functions \( h(t) \):

![Graph showing piecewise-constant feature functions with change points at \( t - L_2 \), \( t - L_1 \), and \( t \).]

\[
p(\mathcal{Y}^u) = \prod_{i=1}^{N^u} \gamma^u(t_i^u | h^u(t_i^u)) \mathbb{1}(t_i^u \in \mathcal{Y}^u) \times \exp\{-\Delta_i^u \gamma^u(t_i^u | h^u(t_i^u))\}.
\]

\( \Delta_i^u = t_{i+1}^u - t_i^u \)

\( N^u \): the number of change points of the feature function \( h^u(t) \).

Figure: Feature construction: \( h^u(t) \)
Assuming piecewise-constant feature functions $h(t)$:

$$p(Y^u) = \prod_{i=1}^{N^u} \gamma^u(t_i^u | h^u(t_i^u)) \mathbb{1}(t_i^u \in \mathcal{Y}^u) \times \exp\{-\Delta_i^u \gamma^u(t_i^u | h^u(t_i^u))\}.$$  

$\Delta_i^u = t_{i+1}^u - t_i^u$  
$N^u$: the number of change points of the feature function $h^u(t)$.  

---  

Figure: Feature construction: $h^u(t)$
Assuming piecewise-constant feature functions $h(t)$:

$$
\prod_{i=1}^{N^u} \gamma^u(t_i^u | h^u(t_i^u)) \mathbb{I}(t_i^u \in \mathcal{Y}^u) \times \exp \{-\Delta_i^u \gamma^u(t_i^u | h^u(t_i^u))\}.
$$

$\Delta_i^u = t_{i+1}^u - t_i^u$

$N^u$: the number of change points of the feature function $h^u(t)$.
Key issue: How to learn the intensity function $\gamma^u(h) : \mathcal{H} \rightarrow \mathbb{R}^+$? Especially in the data-scarce/asynchronized scenarios?
Intensity Function Learning

Key issue: How to learn the intensity function $\gamma^u(h) : \mathcal{H} \to \mathbb{R}^+$?
Especially in the data-scarce/asynchronized scenarios?

- $\gamma(h) = g(f(h)) = (f(h))^2$
Key issue: How to learn the intensity function $\gamma^u(h) : \mathcal{H} \to \mathbb{R}^+$? Especially in the data-scarce/asynchronized scenarios?

- $\gamma(h) = g(f(h)) = (f(h))^2$
- $f(\cdot) \sim GP(m(\cdot), C(\cdot))$
Intensity Function Learning

Key issue: How to learn the intensity function $\gamma^u(h) : \mathcal{H} \to \mathbb{R}^+$? Especially in the data-scarce/asynchronized scenarios?

- $\gamma(h) = g(f(h)) = (f(h))^2$
- $f(\cdot) \sim GP(m(\cdot), C(\cdot))$
- Sharing across $u \rightarrow$ multitask Gaussian processes
Key issue: How to learn the intensity function $\gamma^u(h): \mathcal{H} \to \mathbb{R}^+$? Especially in the data-scarce/asynchronized scenarios?

- $\gamma(h) = g(f(h)) = (f(h))^2$
- $f(\cdot) \sim GP(m(\cdot), C(\cdot))$
- Sharing across $u \rightarrow$ multitask Gaussian processes

\[
\begin{align*}
\mu_0^N & \sim \mathcal{GP}(g, \frac{1}{\xi}K_{NN}), \\
\mathbf{f}_N^u & \sim \mathcal{N}(\mu_0^N, K_{NN}), \\
\gamma_i^u & = \left\{f_{N,i}^u\right\}^2.
\end{align*}
\]

\[
p(Y^u) = \prod_{i=1}^{N^u} (\gamma_i^u)^{\mathbb{I}(t^u_i \in Y^u)} \times \exp(-\Delta_i^u \gamma_i^u)
\]

\[
K_{NN,ij} = \tau^2 \exp \left(-\sum_{p=1}^{P} \frac{(h_{pi} - h_{pj})^2}{2\lambda_p^2}\right) + \sigma^2 \mathbb{I}(i = j)
\]
Multitask Gaussian Processes

Multitask GP:

$$\mu_N^0 \sim \mathcal{GP}\left(g, \frac{1}{\xi}K_{NN}\right)$$  (1)

$$f^u_N \sim \mathcal{N}\left(\mu_N^0, K_{NN}\right)$$  (2)

Multitask linear regression:

$$w^0 \sim \mathcal{N}\left(v, \frac{1}{\xi}I\right)$$  (3)

$$w^u \sim \mathcal{N}\left(w^0, I\right)$$  (4)

$$f^u_{N,i} \sim \mathcal{N}\left(w^u \top h_i, \sigma^2\right)$$  (5)
Multitask Gaussian Processes

Multitask GP:

\[ \mu_N^0 \sim GP \left( g, \frac{1}{\xi} K_{NN} \right) \quad (1) \]
\[ f_N^u \sim \mathcal{N} \left( \mu_N^0, K_{NN} \right) \quad (2) \]

Multitask linear regression:

\[ w^0 \sim \mathcal{N} \left( v, \frac{1}{\xi} I \right) \quad (3) \]
\[ w^u \sim \mathcal{N} \left( w^0, I \right) \quad (4) \]
\[ f_{N,i}^u \sim \mathcal{N} \left( w^u \top h_i, \sigma^2 \right) \quad (5) \]

Choosing a linear kernel for (1) and (2) \( K_{NN,ij} = h_i \top h_j \), and letting \( v = g = 0 \), the above two constructions are equivalent.
Multitask Gaussian Processes

Multitask GP:

\[
\begin{align*}
\mu^0_N & \sim \mathcal{GP} \left( g, \frac{1}{\xi} K_{NN} \right) \\
f^u_N & \sim \mathcal{N} \left( \mu^0_N, K_{NN} \right)
\end{align*}
\]

Multitask linear regression:

\[
\begin{align*}
w^0 & \sim \mathcal{N} \left( v, \frac{1}{\xi} I \right) \\
w^u & \sim \mathcal{N} \left( w^0, I \right) \\
f^u_{N,i} & \sim \mathcal{N} \left( w^u \top h_i, \sigma^2 \right)
\end{align*}
\]

Choosing a linear kernel for (1) and (2) \(K_{NN,ij} = h_i \top h_j\), and letting \(v = g = 0\), the above two constructions are equivalent.

- **Benefit:** choosing more flexible kernels
- **Problem:** \(f^u_N\) and \(\mu^0_N\) are high-dimensional, i.e., \(|\bigcup_u \{h^u_i\}_{i=1}^{N^u}|\) is large.
Two-step construction via a low-dimensional function $f^u_M$:

- $f^u_M$ consists of the function $f^u(\cdot)$ evaluated at $\{s_i\}_{i=1}^M$ (pseudo inputs).
- $\{s_i, f^u_{M,i}\}_{i=1}^M$ captures the characteristics of the function at $\{h^u_i, f^u_{N,i}\}_{i=1}^N$.

\[
\begin{align*}
\mu^0_M & \sim \mathcal{N}\left(g, \frac{1}{\xi}K_{MM}\right) \\
f^u_M & \sim \mathcal{N}\left(\mu^0_M, K_{MM}\right) \\
f^u_N | f^u_M & \sim \mathcal{GP}\left(g + K^u_{NM}K^{-1}_{MM}(f^u_M - g), (1 + \frac{1}{\xi})(K^u_{NN} - K^u_{NM}K^{-1}_{MM}K^u_{NM})^\top\right)
\end{align*}
\]
One-step Construction:

\[ \mu_0^N \sim \mathcal{GP} \left( \mathbf{g}, \frac{1}{\xi} \mathbf{K}_{NN} \right) \]  \hspace{1cm} (2)

\[ \mathbf{f}_N^u \sim \mathcal{N} \left( \mu_0^N, \mathbf{K}_{NN} \right) \]  \hspace{1cm} (3)

Two-step Construction:

\[ \mu_0^M \sim \mathcal{N} \left( \mathbf{g}, \frac{1}{\xi} \mathbf{K}_{MM} \right) \]  \hspace{1cm} (6)

\[ \mathbf{f}_M^u \sim \mathcal{N} \left( \mu_0^M, \mathbf{K}_{MM} \right) \]  \hspace{1cm} (7)

\[ \mathbf{f}_N^u|\mathbf{f}_M^u \sim \mathcal{GP} \left( \mathbf{g} + \mathbf{K}_{NM}^{-1} \mathbf{K}_{MM}^{-1} \left( \mathbf{f}_M^u - \mathbf{g} \right), \right. \\
\left. \left(1 + \frac{1}{\xi} \right) \left( \mathbf{K}_{NN} - \mathbf{K}_{NM} \mathbf{K}_{MM}^{-1} \mathbf{K}_{NM}^{\top} \right) \right) \]  \hspace{1cm} (8)

The marginal prior distributions for \( p(\mathbf{f}_N^u) \) imposed via (2-3) and (7-9) are the same.
Multitask Gaussian Processes: Two-step Construction

One-step Construction:

\[ \mu_N^0 \sim \mathcal{GP} \left( \mathbf{g}, \frac{1}{\xi} \mathbf{K}_{NN} \right) \] (2)

\[ \mathbf{f}_N^u \sim \mathcal{N} \left( \mu_N^0, \mathbf{K}_{NN} \right) \] (3)

Two-step Construction:

\[ \mu_M^0 \sim \mathcal{N} \left( \mathbf{g}, \frac{1}{\xi} \mathbf{K}_{MM} \right) \] (6)

\[ \mathbf{f}_M^u \sim \mathcal{N} \left( \mu_M^0, \mathbf{K}_{MM} \right) \] (7)

\[ \mathbf{f}_N^u | \mathbf{f}_M^u \sim \mathcal{GP} \left( \mathbf{g} + \mathbf{K}_{NM}^{-1} \mathbf{K}_{MM}^{-1} (\mathbf{f}_M^u - \mathbf{g}), \right. \]

\[ \left. (1 + \frac{1}{\xi}) (\mathbf{K}_{NN} - \mathbf{K}_{NM} \mathbf{K}_{MM}^{-1} \mathbf{K}_{NM}^{\top}) \right) \] (8)

The marginal prior distributions for \( p(\mathbf{f}_N^u) \) imposed via (2-3) and (7-9) are the same.

In (6) and (7), \( \mathbf{f}_M^u \in \mathbb{R}^M \), where \( M \ll |\bigcup_u \{ \mathbf{h}_i^u \}_{i=1}^{N_u}| \).
The multitask point process model is summarized as below.

\[ f_u \sim \mu_0, M \overset{g, \xi}{\sim} u_1: N \sim Y_u, \tau, \lambda_1: P, s_1: P \]

**Figure**: Graphical model representation.
Inference Objective

- Model parameters $\Theta$ include the pseudo-input locations, the global mean of transformed rates, and the GP hyperparameters:
  $\Theta = \{\{s_m\}_{m=1}^M, \mu_M^0, \sigma^2, \tau^2, \{\lambda_p\}_{p=1}^P\}$.
- Posterior distribution to learn: $p(f_M^u, f_N^u | \mathcal{Y}, \Theta)$.

\[
p(f_M^u, f_N^u | \mathcal{Y}, \Theta) \approx q(f_N^u, f_M^u) \quad (9)
\]
\[
= p(f_N^u | f_M^u) q(f_M^u) \\
= p(f_N^u | f_M^u) \mathcal{N}(f_M^u; \mu^u, \Sigma^u) \quad (10)
\]
Model parameters $\Theta$ include the pseudo-input locations, the global mean of transformed rates, and the GP hyperparameters:

$$\Theta = \{ \{ s_m \}_{m=1}^M, \mu_0^M, \sigma^2, \tau^2, \{ \lambda_p \}_{p=1}^P \}.$$

Posterior distribution to learn: $p(f_M^u, f_N^u | \mathcal{Y}, \Theta)$.

$$p(f_M^u, f_N^u | \mathcal{Y}, \Theta) \approx q(f_N^u, f_M^u) \approx p(f_N^u | f_M^u)q(f_M^u) = p(f_N^u | f_M^u)N(f_M^u; \mu^u, \Sigma^u).$$ (10)

Using Variational EM

- Variational E-step, fix $\Theta$ is fixed and optimize $\{ \mu^u, \Sigma^u \}_{u=1}^U$
- Variational M-step, fix $\{ \mu^u, \Sigma^u \}_{u=1}^U$ and optimize $\Theta$. 
Data Source: New Zealand national minimum dataset, 2007-2011
Cleaned Dataset: visit streams of multiple patients for each disease category
- 6 chronical disease categories
  - e.g., neoplasms, metabolic problems (including type-I and type-II diabetes)
- Number of patients: 36 – 118
- Number of visits per patient: 51 – 98
- split visit sequences for each patient into training and testing (half split)
EHR Dataset: Data Description

Data Source: New Zealand national minimum dataset, 2007-2011
Cleaned Dataset: visit streams of multiple patients for each disease category
- 6 chronic disease categories
  - e.g., neoplasms, metabolic problems (including type-I and type-II diabetes)
- Number of patients: 36 – 118
- Number of visits per patient: 51 – 98
- split visit sequences for each patient into training and testing (half split)

Setup
- Features: number of visits in previous week, month, and 6 months
- Number of pseudo inputs: $M = 15$
- Task: predicting the arrival pattern in $[T, T + L]$, e.g., $L = 30$
Figure: Model fit results. Top-left: Comparison of data log-likelihood per day of MTPP, direct inference, and thinning approach; Top-right and bottom row: Intensity functions inferred of anonymous patients’ arrival sequences.
### Table: AUC of binary predictions of events occurring in weekly and monthly windows for 6 disease types.

<table>
<thead>
<tr>
<th>Disease Type</th>
<th>1 Week</th>
<th>1 Month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MTPP</td>
<td>PoiR</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>0.7379</td>
<td>0.7249</td>
</tr>
<tr>
<td>Metabolic</td>
<td>0.6807</td>
<td>0.6170</td>
</tr>
<tr>
<td>Nervous</td>
<td>0.6926</td>
<td>0.7241</td>
</tr>
<tr>
<td>Circulatory</td>
<td>0.6807</td>
<td>0.6778</td>
</tr>
<tr>
<td>Respiratory</td>
<td>0.5733</td>
<td>0.6302</td>
</tr>
<tr>
<td>Digestive</td>
<td>0.6050</td>
<td>0.5562</td>
</tr>
</tbody>
</table>
Figure: Neoplasms results. Global rate function inferred by MTPP as a function of the number of arrivals in windows \([t - 7, t]\) and \([t - 30, t]\).
Summary

- Analyzed multiple streaming point processes in a multi-task setting
- Proposed a simple strategy for forward prediction
- Constructed hierarchical GPs to leverage information across the tasks
Acknowledgements

Thanks for helpful discussions with
- Christopher Meek, Microsoft Research Redmond
- Asela Gunawardana, Microsoft Research Redmond

And fundings from
- ARO, DARPA, DOE, NGA and ONR.


