Mining Medical Records with a KLIPI Multi-Dimensional Hawkes Model

Yinan Zhao*
ZYN257@sjtu.edu.cn

Xiaojuan Qi*
321456@sjtu.edu.cn

Zhengzhe Liu
liuzhengzhelzz@gmail.com

Ya Zhang†
ya_zhang@sjtu.edu.cn

Tao Zheng‡
18918119818@126.com

ABSTRACT

How does the disease evolve? What is the relationship between different diseases? A model describing the underlying disease evolvement process and disclosing the interaction between disease is significant for health promotion, disease prediction, disease relationship discovery and so on. In this paper, we proposed the Kernel Learning with Individual Physique Indicators(KLIPI) multi-dimensional Hawkes model. This model captures both the natural incidence of disease and the triggering effect of the past medical history of the patients. We use Gaussian density estimator to acquire the triggering kernel accurately. What's more, to represent the physique variations among patients, we include the Individual Physique Indicators(IPI) into our framework. We validate our model with experiments on medical datasets. The experiments show evidently that our model outperforms the original multi-dimensional Hawkes model and the Markov model.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical sciences-Medical information systems

General Terms

Algorithms,Design

Keywords

KLIPI Hawkes Model,medical records,MM algorithm,kernel learing,physique

1. INTRODUCTION

Medical data mining has gained popularity with the development of computer aided medical system which enables people to observe the medical data directly. Consequently, many efforts were made to disease prediction, disease relationship discovery and so on. Disease prediction based on patients' symptom was investigated in[6][8], which predict the possibility of getting disease based on the input features, e.g. blood pressure, age, sex and so on. But these models have a drawback that they just make prediction based on the current symptoms of the patients, and hence can't predict the development of the disease in the long run. Decision making model based on temporal sequence of patients was studied in[1]. They simulate the disease treatment process as a MDP and take the history of the patients into consideration, but they require time to be discrete and hence can't take diseases in different time scale into their frame of work. In[4],Jose discussed the application of HMM model on disease relationship discovery and they treat time as a continuous variable but they only consider one-order Markov process so they can't capture the connection of diseases which are not consecutively related. All the above models have a common deficiency that they can't be interpreted in the medical point of view.

Modeling the disease evolvement process from medical point of view is essential for disease prediction and disease prevention. What's more, they can also promote the development of basic medical research and help people understand the disease better. But so far little effort has been made in modeling the disease evolvement process medically. We propose a model to describe the disease evolvement process and reveal the relationship between disease. We focus on the Electronic Health Record(EHR). Each record contains the information of one clinical visit, i.e. patient ID, sex,age,time of clinical visit, disease. All the diseases are shown of their corresponding ICD10 code. We show the example record in Table 1.

We rearranged the records according to the patient ID. Clinical visits of the same patient produce a temporal sequence of disease. These sequence can reveal the latent pattern of disease evolvement. For example if two diseases occurred on the same patient at different time and the co-occurrence happened on a statistically large number of patients, then

---

** means equal contribution.

†Ya Zhang, Yinan Zhao, Xiaojuan Qi and Zhengzhe Liu are with the Institute of Image Communication and Network Engineering & Shanghai Key Laboratory of Multimedia Processing and Transmissions, Shanghai Jiao Tong University, Dongchuan RD.800, Shanghai 200240,China. Ya Zhang is the corresponding author.

‡Tao Zheng is with Changning Health Information Center, Shanghai.

Table 1: EXAMPLE RECORD

ID	Sex	Age	Time	Disease
808210	female	59	2009-1-1	I25.101
526855	female	51	2009-1-1	J06.903
765972	male	61	2009-1-2	E14.901

a causal connection between the two diseases can be discovered[4]. We study the time sequence to uncover the underlying process. We assume that there are three factors influencing the possibility a person get a disease, e.g. the natural incidence of the disease, the individual physique, and the past medical history.

In this paper, we proposed a Kernel Learning with Individual Physique multi-dimensional Hawkes model(KLIPI Hawkes model) to discover the underlying disease evolvement process. The multi-dimensional Hawkes model captures the base intensity effect and the mutually-triggering effect ,which correspond to the natural incidence of disease and the past medical history of the individual. The model is designed as a generalization one, so we will not take some factors like genetics, environment influence into account.Inspired by [9], we also take the individual physique into our framework by introducing the IPI(Individual Physique Indicators) into the model. In order to make the model describe the underlying process more accurately, we creatively use the gaussian kernel estimator to learn the triggering kernel of the model. We experiment with the medical dataset to validate the effectiveness of our method and compare with Markov model in disease incidence prediction.

The paper is organized as follows. In section 2, we briefly describe the original multi-Hawkes model. In section 3,the proposed model and algorithm is described in detail and we introduce IPI and kernel learning into our model. In section 4, we give a brief description of the dataset. In section 5, we experiment with medical dataset to verify the effectiveness of our method and improvement in describing the disease evolvement process.

2. HAWKES MODEL

We design our KLIPI model mainly based on the principle of Hawkes model which is often used for modelling social infectivity in social network[10][9][11]. So we will start from introducing the basic form of one-dimensional Hawkes model in this section.

2.1 One-dimensional Hawkes model

We firstly describe one-dimensional Hawkes process briefly. In the next subsection we will describe multi-dimensional Hawkes model applied in our medical record modelling in detail. A basic one-dimensional Hawkes process is a special point process[7]with its conditional intensity function shown as follows.

$$\lambda^*(t) = \mu(t) + \alpha \sum_{t_i<t} \gamma(t - t_i; \beta) \tag{1}$$

where $\mu(t) \geq 0$ denotes the base intensity, meaning the natural intensity which is not the reflection of events happened before. $\alpha > 0$ describes the intensity of self-exciting nature. Larger α indicates events in the past time have a more positive contribution to the events in the future. t_i are the time of events in the process before time t, and $\gamma(t; \beta)$ is the self-exciting decay kernel function. We will focus on exponential kernel expressed as follows to discuss the self-exciting nature of one-dimensional hawkes model.

$$\lambda^*(t) = \mu(t) + \alpha \sum_{t_i<t} \beta exp(-\beta(t - ti)) \tag{2}$$

When an event happens , it will increase the probability of other events in a short time, and approach approximately to μ in a longer time under the decaying process.

2.2 Multi-Dimensional Hawkes model

In order to model the disease correlation, we extend one-dimensional Hawkes model to multi-dimensional condition, which is composed of different one-dimensional hawkes processes. In this paper, we use U to denote the dimension of multi-dimensional hawkes model, and the model is a process of U-dimension denoted as N_t^u,$u = 1, ..., U$. The conditional intensity function for the u-dimensional Hawkes model is expressed as follows.

$$\lambda_u(t) = \mu_u + \sum_{i:t_i<t} a_{uu_i} \gamma(t - t_i) \tag{3}$$

where $\mu_u \geq 0$ is the base intensity for the u-th Hawkes process. The coefficient a_{uu_i} captures the mutually-exciting property between the u-th and u_i-th dimension. Intuitively, it captures the degree of influence of events occurred in the u-th dimension on the u_i-th dimension. Larger value of a_{uu_i} indicates a larger correlation between these two dimensions and events in the u-th dimension are more likely to trigger events in the u_i-th dimension in the future. For further discussion, we collect the parameters into matrix, denoted as $\boldsymbol{\mu} = (\mu_u)$ for the base intensity of each dimension and $\boldsymbol{A} = (a_{uu_i})$ for the mutually-exciting coefficients. We require both $\boldsymbol{A}$ and $\boldsymbol{\mu}$ to be nonnegative.

3. KLIPI MODEL

3.1 Model Description

As we believe some past-time diseases may have positive influence on the disease occurred on a patient in the future, so we adopt the multi-dimensional Hawkes model. We mainly focus on the influence of past-time diseases on one individual, consequently, contagion between different individuals is not our consideration. Further more, as people of different degree in physique indicator show various resistance to the diseases, we bring in the concept of individual physique indicator. In addition, many works based on Hawkes model choose exponential function as the kernel and preset β, for better fitting the medical data, we also improve our model with Gaussian kernel density estimation for kernel learning. In this section, we will illustrate our modified model in detail.

3.1.1 Modified Model with Physique

As we all know, difference of individual's physique is the basic reason why different individuals have various resistance to diseases. Based on this idea, we introduce b_c to describe the degree of individual's physique into the basic multi-dimensional Hawkes model, and value of b_c will change the degree of conditional intensity.

With physique b_c, the conditional intensity function appears different to various individuals. The basic form of the improved Hawkes model is expressed as follows. Larger b_c indicates the higher degree of the conditional intensity, which means in the same circumstance a patient is more likely to suffer from diseases with smaller degree of resistance.

$$\lambda_u^c(t) = b_c(\mu_u + \sum_{i:t_i<t} a_{u^c u_i^c} \gamma(t - t_i^c)) \tag{4}$$

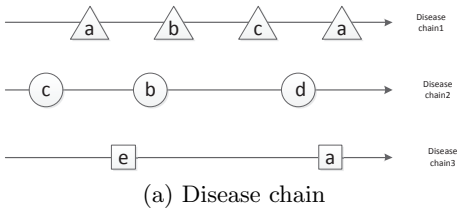
According to the parameters discussed in the last section, we illustrate the meaning of each parameter in our model as follows.

- $u(u = 1, 2, ..., U)$ is the dimension of Hawkes model, corresponding to some kind of disease in our model like diabete.
- $\lambda_u^c(t)$ is the conditional intensity function of the c-th sample and u-th event. In our model, it denotes the intensity of disease at time t for patient c, which indicates larger $\lambda_u^c(t)$ leads to bigger risk of disease at time point t for patient c. We can get the probability of the u-th disease at some time in the future by calculating $\lambda_u^c(t)$.
- μ_u is the base conditional intensity function, also called natural intensity of the event of the u-th dimension. In our model, it represents the statistical risk of the diseases.
- a_{uu_i} is the coefficient of influence intensity of u_i-th event on u-th event. In our proposed model, it corresponds to the intensity of mutually-exciting nature between the u_i-th disease and the u-th disease. In other words, larger value of a_{uu_i} indicates stronger correlation of the two diseases, which may explain a higher risk of having u-th disease in the future when a patient already suffers from the u_i-th disease in the past time.
- $\gamma(t)$ is the decay kernel of the influence between diseases. In our model, the impact of one disease on another change with the time.
- b_c represents the Individual Physique Indicator(IPI). Larger b_c indicates less resistance to diseases.

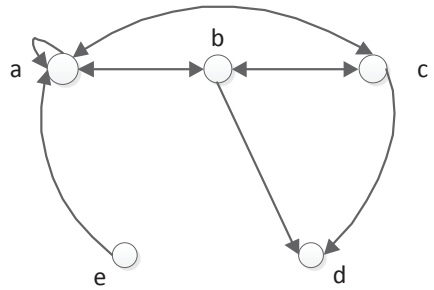
The graphical illustration of multi-dimensional Hawkes model is shown in Figure 1. Figure (a) represents the real disease chain of patients, and a,b,c,d,e represent real diseases. By modelling the observed records, figure (b) reveals the inner correlation of every disease. The arrow corresponds to $\alpha_{u_i u_j}$ in the model, and each circle indicates the base intensity μ_u. The size of circles shows the degree of base intensity of each kind of disease.

3.1.2 *Kernel Learning*

In previous works the kernel function is mainly empirically set as exponential function $\gamma(t) = \beta exp(-\beta t)$, while some real diseases may not influence others in that assumed way. We use the non-parametric kernel estimator to realize kernel learning process. As an important limitation of the parametric kernel estimator is that the chosen density may be



(a) Disease chain



(b) Underlying network

Figure 1: Disease chain and latent correlation. The lager circle means higher natural incidence, and the arrows means the triggering effect.

a poor approximation of the distribution that generates the data, which may lead to inaccurate predictive performance, we use Gaussian kernel density estimator for kernel learning in our KLIPI model.

Here we will give a brief introduction of Gaussian kernel density. As we assume the kernel function to be a Gaussian distribution in this method, when we get the training data sample $A = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ...\boldsymbol{x_N}\}$, we can get the estimation probability density function as follows.

$$f(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{1/2}} exp\bigg(- \frac{||\boldsymbol{x} - \boldsymbol{x}_n||^2}{2h^2} \bigg) \tag{5}$$

where h is pre-set sensitive value. The parameter h is a smoothing parameter, and there is a trade-off between sensitivity to noise when h is too small and over smoothing when h gets too large. So the parameter h needs to be moderate for kernel density estimation.

3.2 Model Parameters Estimation

For the estimation process, suppose we have m chains of medical records which are listed in time-sequential order, denoted as $\{c_1, c_2, ..., c_m\}$, the length of the chains denoted as n_c, so each medical record can be written as $\{({t_i}^c, {u_i}^c)\}_{i=1}^{n_c}$. The notations used in this section are listed in Table 2. According to the principle of Maximum Likelihood Estimation, we adopt MM algorithm which is mainly used for optimization of high-dimension problems to optimize the value of likelihood function[2]. We start the optimization process by calculating the likelihood function which is described as follows. The parameters are denoted as $\Theta = \{\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{b_c}\}$

$$\begin{aligned} l(\Theta) = \sum_c \bigg(\sum_i^{n_c} log(\lambda_{u_i^c}^c(t_i^c)) - \big(T_c \sum_u \mu_u \\ + \sum_{u=1}^{U} \sum_{j=1}^{n_c} a_{u u_j^c} \Gamma(T_c - t_j^c)\big) \bigg) \end{aligned} \tag{6}$$

Table 2: Notations used in estimation

Symbol	Description
c	Time chain of patient c
n_c	The length of chanin c
u	u-th disease
μ_u	base intensity of the u-th disease
T_c	time length of time chain c
$\lambda_u^c(t)$	intensity of u-th disease for patient c
$\alpha_{u_i u_j}$	degree of influence of the u_j-th disease on the u_i-th disease
$\gamma(.)$	decay kernel function of influence between diseases

where $\Gamma(t) = \int_0^t \gamma(s)ds$.

The parameters can be estimated by maximizing the log-likelihood function $max_{\boldsymbol{A}\geq 0, \boldsymbol{\mu}\geq 0} l(\Theta)$. We will show the optimization process of KLIPI model using MM algorithm in detail. Firstly, for the minorize process we calculate the lower bound of the likelihood function shown as follows.

$$\begin{aligned} l(\Theta) &= \sum_c \Bigg(\sum_i^{n^c} log\big(\mu_{u_i^c} + \sum_{j:t_j^c<t_i^c} a_{u_i^c u_j^c}\gamma(t_i^c - t_j^c)\big) \\ &\quad - \big(T_c \sum_u \mu_u + \sum_{u=1}^{U}\sum_{j=1}^{n^c} a_{uu_j^c}\Gamma(T_c - t_j^c)\big)\Bigg) \\ &\geq \sum_c \Bigg(\sum_i^{n^c} \big(p_{ii}^c log\frac{\mu_{u_i^c}}{p_{ii}^c} + \sum_{j=1}^{i-1} p_{ij}^c log\frac{a_{u_i^c u_j^c}\gamma(t_i^c - t_j^c)}{p_{ij}^c}\big)\Bigg) \\ &\quad - \sum_c \Bigg(T_c \sum_u \mu_u + \sum_{u=1}^{U}\sum_{j=1}^{n^c} a_{uu_j^c}\Gamma(T_c - t_j^c)\Bigg) \\ &= Q(\Theta|\Theta^{(k)}) \end{aligned} \tag{7}$$

where

$$\begin{aligned} p_{ij}^c &= \frac{a_{u_i^c u_j^c}^{(k-1)}\gamma(t_i^c - t_j^c)}{\mu_{u_i^c}^{(k-1)} + \sum_{j=1}^{i-1} a_{u_i^c u_j^c}^{(k-1)}\gamma(t_i^c - t_j^c)}, j = 1, ..., i-1 \\ p_{ii}^c &= \frac{\mu_{u_i^c}^{(k-1)}}{\mu_{u_i^c}^{(k-1)} + \sum_{j=1}^{i-1} a_{u_i^c u_j^c}^{(k-1)}\gamma(t_i^c - t_j^c)} \end{aligned} \tag{8}$$

We verify the lower bound $Q(\Theta|\Theta^{(k)})$ meets both two conditions of MM algorithm by Jensen's inequality expressed as follows.

$$\begin{aligned} l(\Theta) &\geq Q(\Theta|\Theta^{(k)}), \forall\Theta \\ l(\Theta^{(k)}) &= Q(\Theta^{(k)}|\Theta^{(k)}) \end{aligned} \tag{9}$$

For the maximizing process, the value of Θ is calculated for every time of iteration. The maximum value can be directly calculated through derivation, and the iteration results are listed as follows.

$$\begin{aligned} \mu_u^{(k)} &= \sqrt{\frac{\sum_c \sum_{i:i\leq n^c, u_i^c=u} p_{ii}^c}{\sum_c b_c^{(k-1)} T_c}\mu_u^{(k-1)}} \\ a_{uu'}^{(k)} &= \sqrt{\frac{\sum_c \sum_{i:u_i^c=u} \sum_{j:j<i, u_j^c=u'} p_{ij}^c}{\sum_c \sum_{j:u_j^c=u'} b_c^{(k-1)}(G(T_c - t_j^c) - G(0))}a_{uu'}^{(k-1)}} \\ b_c^{(k)} &= \sqrt{\frac{n_c}{T_c \sum_u \mu_u^{(k-1)} + \sum\sum_{j=1}^{n^c} a_{uu_j^c}^{(k-1)}(G(T - t_j^c) - G(0))}b_c^{(k-1)}} \end{aligned} \tag{10}$$

The iteration process continues till reaching the convergence when the value of log-likelihood does not change any more.

4. MEDICAL DATASET

We evaluate the proposed Kernel Learning with Individual Physique(KLIPI) multi-dimensional Hawkes Model on the data set provided by Changning Health Center in Shanghai, China. One medical visit with several kind of disease is handled as several medical visits. It contains 1034816 patients' medical records from January 1, 2009 to December 31, 2013 and 5721 kind of diseases. Here we restrict the dataset to patients with at least 2 medical visits. In order to avoid underfit, we consider the diseases with more than 10000 records in our experiments. The total kinds of disease studied is 125.We illustrate these numerically data in Table 3.

Table 3: DATASET

Property of Dataset	Concrete Statistic
Number of patients	1034816
Starting time point	2009-01-01
Ending time point	2013-12-31
Kind of disease	5721
Kind of target disease	125

The statistical results for number of patients and number of clinical visits are shown in Figure 2. The x-axis shows number of clinical visits, and the y-axis shows number of patients who have exact x clinical visits.

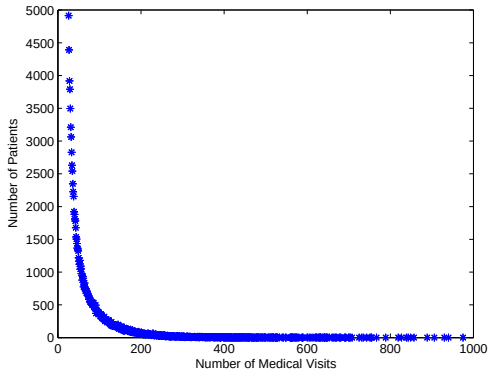


Figure 2: Medical Visits of Patients

5. EXPERIMENTS

In this part, we experiment with the medical dataset to evaluate the performance of our model. And then we compare the proposed KLIPI multi-dimensional Hawkes model with both the original multi-dimensional Hawkes model and Markov model based on the Medical Dataset.

5.1 Model Convergence

Here we demonstrate the convergence of our proposed method using the medical dataset. We use the MM algorithm iteratively to maximize the Log-likelihood function. When the increase of the Log-likelihood between two consecutive iterations is less than 0.02% of the former negative Log-likelihood, the Log-likelihood is considered to be stable and the algorithm reaches convergence. It is shown in Figure 3. The figure shows that with the number of iterations increases, the Log-likelihood increases monotonically, and the iteration doesn't stop until the convergence condition is satisfied.

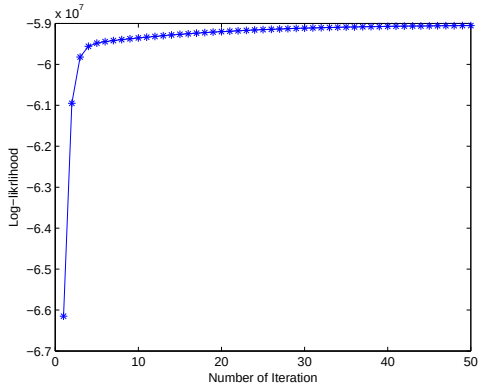


Figure 3: Log-likelihood

5.2 Model Interpretation

In the previous section, we interpret the model parameter in medical view. In this section we validate our interpretation with the medical dataset.

5.2.1 *Parameter b_c and Individual Physique*

In the previous section, we show that parameter b_c is related with the individual's physique. We will get lower b_c from people who have a stronger physique. Assume that the patients' physique is anti-related to the number of medical visits in the five years of records. The assumption is sensible since people with weaker body will be more vulnerable to different kinds of disease, and they will have more clinical visits during the same time period. So we use the number of medical visits to represent the physique of the patient. In other words, to verify our interpretation of b_c, we need to show that b_c has positive correlation with the number of clinical visits. Since b_c and the number of clinical visits is not numerically comparable, we use the Spearman Correlation Coefficient between b_c and the number of clinical visits to represent the accuracy of b_c in predicting the individual Physique.In our dataset the Spearman Correlation Coefficient is 0.9013 which shows strong relation between our parameter b_c and the individual physique. Therefore, the result matches our interpretation of b_c excellently. We use least square regression to analysis 100 patients, and the rank of clinical visits and the rank of b_c are shown in Figure 4.

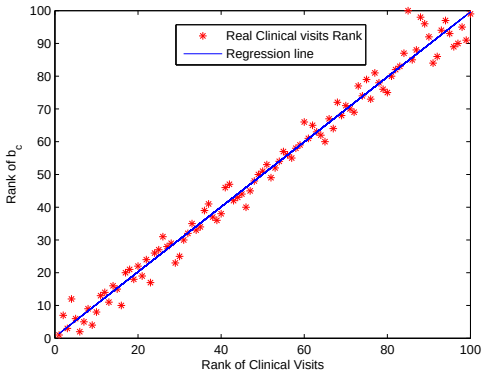


Figure 4: Least Square Regression of clinical visits and b_c by ranking

5.2.2 *Case study*

In this section, we mainly focus on the results of matrix $\boldsymbol{A}$. As we have discussed, $\boldsymbol{A}$ capture the intensity of mutually-exciting nature between two diseases which indicates larger $a_{uu'}$ is a reflection of stronger correlation between diseases. These may explain a higher risk of having some disease in the future when a patient already suffers from the correlated disease in the past time. For better illustration, we calculate the average value of the elements in $\boldsymbol{A}$ and choose some elements among which are much larger than the average value. In this section, we will both interpret the model qualitatively and quantitatively. We will choose some correlations between diseases which may be obvious in clinical field for better interpretation of the model efficiency qualitatively. The correlations are listed in table as follows.

Table 4: Diseases Correlation

prior disease	latter disease
hypertension	diabetes
cataract	conjunctivitis
hypertension **I**	coronary disease
hepatitis B	hypertension
coronary stenosis	hyperlipemia
uremia	hypertension **II**
benign hypertension	benign hypertension

We will illustrate the table in three aspects to explain the correlation more clearly. Firstly, we can get some direct correlation between diseases, like hypertension and diabetes which accords with our common sense. The direct correlation explains some co-existence diseases happening to the patients. Secondly, we can get some indirect pairs of diseases in different time. Take hepatitis B and hypertension for instance. As we know that for the patient who suffers from hepatitis B there is often a disorder of glucose which may

lead to diabetes. Combining with our common knowledge, we can conclude there is indirect correlation between these two diseases. Furthermore, some coronary diseases and hypertension may form a circle of correlation. Finally, some coefficients show the self-exciting nature of some diseases which indicates worse condition of the patients.

5.2.3 *Parameter μ_u and Disease Incidence*

We have shown that the parameter $\mu_u (u = 1, 2, ...U)$ represents the base disease incidence and is positive related to the overall disease incidence of our dataset. We get the overall disease incidence by counting the frequency of the studied disease in our dataset and then normalize the result. Then we compare the result with the base disease incidence predicted by parameter μ_u which also needs normalization. We use least square regression to visualize the correlation between the normalized μ_u and overall disease incidence.The result is shown in Figure 5. The x-axis represents the overall

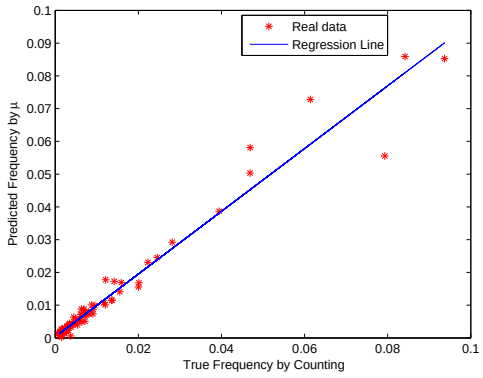


Figure 5: Interpretation of Parameter μ

disease incidence $I_u (u = 1, 2, ...U)$, and the y-axis represents the normalized $\mu_u (u = 1, 2, ...U)$. The predicted μ_u increase linearly with I_u, and the correlation between the normalized μ_u and the overall disease incidence I_u is 0.9579. μ_u and I_u show strong linear correlation, which verify our assumption that mu_u represents the base disease intensity.

5.2.4 *Estimated Kernel*

We use Gaussian Kernel density estimator to learn the Triggering Kernel function. Triggering Kernel function describes the temporal variation of the triggering effect. If one disease has a remarkable triggering effect on another disease at time t , the possibility of metabasis will be high. In the dataset, many patients develop one disease at prior time and get another disease after a time interval t . So we estimate the kernel function based on the number of patients who have disease metabasis at time t. In this experiment the minimum time interval is set as one day. The time interval between two diseases is represented by the days between the two diseases.The longest interval of metabasis between two disease is 1822 days. We count the number of metabasis of diseases at time interval ranging from 0 to 1822. We use a Gaussian function to estimate the pattern of metabasis of disease as discussed before. Here we set the standard deviation $h = 10$ based on the distribution of the time interval. The estimated kernel function and distribution of the real data are shown in Figure 6. The x-axis represents time, and the y-axis represents the intensity of metabasis between diseases. The figure shows the estimated kernel matches the data pattern accurately. It captures the variation of metabasis intensity with time.

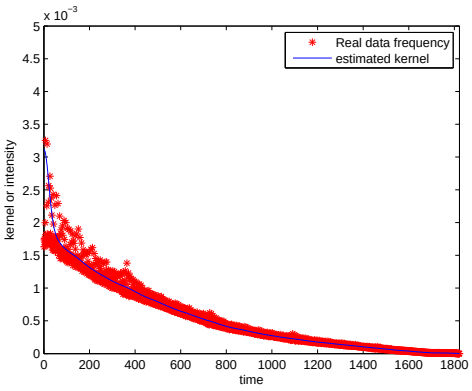


Figure 6: Estimated Kernel

5.3 Accuracy Comparison and Analysis

In this section, we compare the proposed modification of the model with the original model to show that these improvements describe the evolvement of diseases more accurately. Then we compare the proposed method with the Markov model method in the prediction of disease incidence. In this process, we split the medical data into two sets: the training set and the testing set. To guarantee the objective and sufficient of the comparison ,we design four hand-out validation tests. The training set contains 50%, 60%, 70% and 80% records of every time sequence respectively, and the rest are separately used as test data to verify the model accuracy in prediction. We illustrate the number of clinical visits in the training and testing set in Table 5.

Table 5: Splitting Dataset

Splitting percent	Training set	Testing set
50%	6502753	6935220
60%	7657573	5780400
70%	8969991	4467982
80%	10309166	3128807

The method is as follows:

- Use the training dataset to fit the model .In this step we get the model parameters Θ;
- Caculate the conditional intensity of each kind of disease based on the testing time points using the model parameters.In this step we get the predicted disease incidence intensity ;
- Calculate the Spearman Correlation Coefficient between the predicted disease incidence intensity and the statistical incidence of the testing datasets. In this step we get the prediction accuracy of the model as well.

5.3.1 The Effect of Individual Physique Index

We evaluate the effect of Individual Physique Index b_c by comparing the prediction accuracy between Models with b_c and models without b_c. To guarantee the objectivity of the evaluation, we fix Kernel function as $\gamma(t) = \beta exp(-\beta t)$. We set $\beta = 0.005, \beta = 0.05$ and $\beta = 0.5$ respectively, and train the models with $50\%, 60\%, 70\%$ and 80% records of the time sequence. The result of models with b_c and without b_c are shown in Figure 7.

Three subfigures($\beta = 0.005, \beta = 0.05, \beta = 0.5$)show the comparison of accuracy between models with Individual Physical Index b_c and models without b_c, the white block correspond to the models without Individual Physical Index(IPI) and the blue block correspond to the models with IPI. It is evident that the models with IPI performs better than models without IPI in prediction accuracy. Although in Figure 7(c) the improvement is not quite significant, the overall prediction accuracy increases especially for the models with kernel parameter β=0.005. Introducing physique into the original Hawkes model, we make the model more accurate in prediction of disease incidence .

5.3.2 The Effect of Kernel Learning

In this part, we verify the effect of kernel learning by comparing the prediction accuracy of models with fixed kernel and models with kernel learning. We consider the effect of IPI in both models . For the fixed kernel model, we use the prevalent kernel $\gamma(t) = \beta exp(-\beta t)$, and preset $\beta = 0.005, \beta = 0.05, \beta = 0.5$. Similarly, we use the four different training set to train our model and the rest of the training set to testify the prediction accuracy. The result is shown in Figure 8.

Three subfigures are named according to the fixed kernel parameter($\beta = 0.005, \beta = 0.05, \beta = 0.5$). The white block shows the prediction accuracy of the fixed kernel and the blue block shows the accuracy of the learned kernel model. Our method increase the prediction accuracy significantly. The model with kernel learning outperforms the model with fixed kernel most of the time except one condition when $\beta = 0.005$ and the training percent is 50%. The exception may caused by the randomness of the algorithm. Figure 9 shows the variation of prediction accuracy over the training set percent. The prediction accuracy of the Kernel Learning model increase with the percentage of the training set and stay stable when the training set is large enough. But the model with fixed kernel show a decrease in prediction accuracy when the testing set becomes small enough. This demonstrates that the fixed kernel method is vunerable to noise. When the testing set is small, it is easily influenced by the noise in the data and can't behave as good as before. But our kernel learning method is not influenced by the small testing set and behaves as good as before. So in this respect, the kernel learning method can be more resistable to noise comparing to the fixed kernel method.

5.3.3 Markov Model Comparison

In this part, we compare our method (KLIPI multi-dimensional Hawkes model) with the one-order Markov model in prediction accuracy. One-order Markov model is widely used in prediction based on time sequence and is frequently used

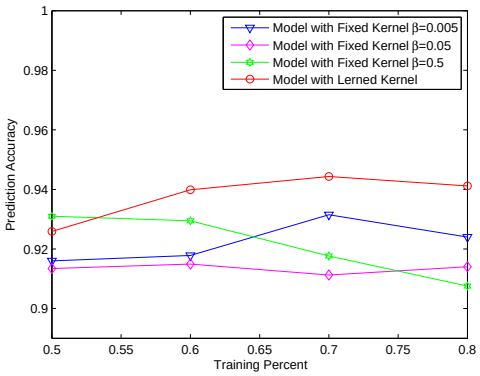


Figure 9: Prediction Accuracy with Training Percent

in disease prediction[5][4][3] and disease relationship discovery[4]. Here we compare our model with the one-order Markov model.The result is shown in Figure 10.

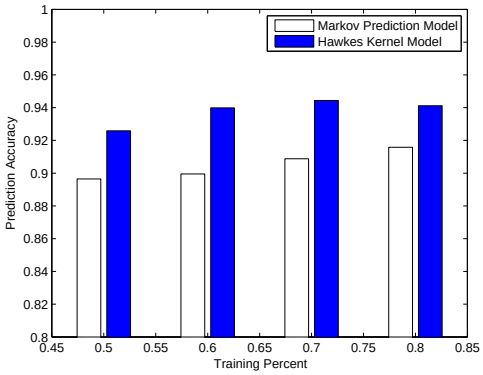
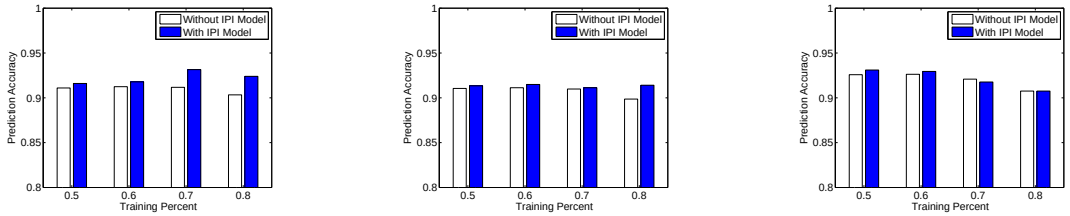


Figure 10: Comparison between Markov Model and Hawkes Kernel model

The white block in the figure shows the Markov model, and the blue block shows our model in prediction of the disease incidence . It's obvious that our method outperforms the Markov prediction method. Figure 11 shows the prediction variation with the training percent. It shows that when the training set is small,the prediction accuracy of the Markov model is much lower. When the training set gets larger, which also means the testing set becomes smaller, Markov model behaves more accurately. This means that the Markov model behaves bad in predicting over a long time. While our method performs always good even when it predicts over a large time scale. In this respective, Our method can predict over a long time, and overcomes the drawback of the Markov model.

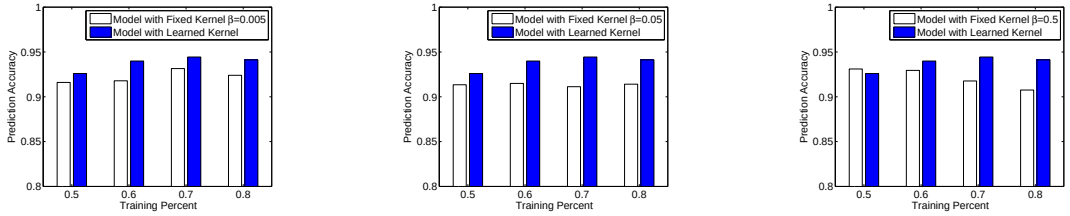
6. CONCLUSIONS

In this paper, we propose a improved KLIPI model based on multi-dimensional Hawkes model. The proposed model indicates the correlation between different kinds of diseases and disease evolvement as well. We introduce a parameter representing individual physique and use Gaussian kernel density



(a) $\beta = 0.005$ (b) $\beta = 0.05$ (c) $\beta = 0.5$

Figure 7: Accuracy Comparison between model with b_c and without b_c.



(a) $\beta = 0.005$ (b) $\beta = 0.05$ (c) $\beta = 0.5$

Figure 8: Accuracy Comparison between model with Kernel Learning and without kernel Learning.

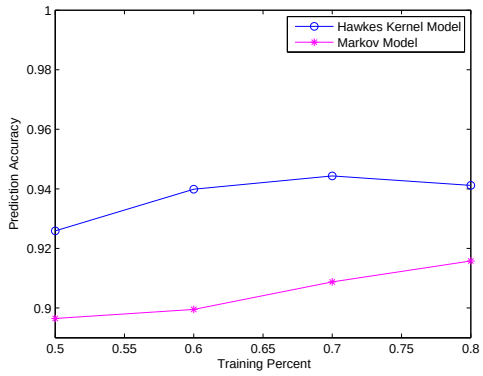


Figure 11: Prediction accuracy with Training percent(Markov and Our method)

estimator for the estimation of the kernel function. The modifications both increase the accuracy of prediction on disease evolvement and we also prove KLIPI Hawkes model outperforms Markov in disease prediction.

There are several promising directions for our study in the future. Firstly, we plan to take the initial disease time of patients into account. Secondly, we will investigate the treatment information to realize the integrated model. In particular, we plan to set up the supporting decision system for doctors according to the accurate prediction.

7. ACKNOWLEDGMENTS

This work is supported by the High Technology Research and Development Program of China (2013AA020418).

8. REFERENCES

[1] J. R. Beck and S. G. Pauker. The markov process in medical prognosis. *Med Decis Making*, 3(4):419–458, 1983.

[2] Hunter, D. R, and K. Lange. A tutorial on mm algorithms. *The American Statistician*, 56(1):30–37.

[3] Mitchell and C. E. *Semi-Markov Multi-state Modeling of Human Papillomavirus*. PhD thesis, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL, USA, MAY 2012.

[4] J. M.Leiva-Murillo. Visualization and Prediction of Disease Interactions with Continuous-time Hidden Markov Models.

[5] H. O.Geman. Parkinson's disease prediction based on multistate markov models. *INT J COMPUT COMMUN*, 8(4):525–537, 2013.

[6] Pattekari, A. Shadab, and A. Parveen. Prediction System For Heart Disease Using Naive Bayes. *International Journal of Advanced Computer and Mathematical Sciences ISSN*, pages 2230–9624.

[7] J. G. Rasmussen. Temporal point processes: the conditional intensity function. 2009.

[8] Sudha and S. Disease Prediction in Data Mining Technique–A Survey. *IJCAIT*, 2(1):17–21, 2013.

[9] Y. Wei, K. Zhou, Y. Zhang, and H. Zha. Learning the hotness of information diffusions with multi-dimensional hawkes processes. *Agents and Data Mining Interaction*, pages 92–110, 2014.

[10] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 641–649, 2013.

[11] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional hawkes processes. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1301–1309, 2013.