論 文 の 内 容 の 要 旨

論文題目　Variational Bayesian Inference of Point Processes for Time-Sequence Modeling
（時系列モデリングのための点過程の変分ベイズ推論）

氏　名　ディン　ホンイ
丁　弘毅

A time-sequence consists of a set of time-stamps, each of which records the arrival time of an event. Time-sequence data can generally be classified into two types. One is from experiments that monitor subjects in a continuous fashion; and thereby the exact timestamps of all occurrences of the events are fully observable. These data are usually referred to as recurrent event data. On the other hand, we have the so-called panel count data, in which only the numbers of occurrences of the events between subsequent observation times. In real-world problems arising in areas such as social science, health care and crime prevention, time-sequence modeling is extremely useful since it can help us in predicting future events and understanding the reasons behind them.

A common approach to time-sequence modeling is to assume a time-sequence is generated by a temporal point process. Cox processes are widely used in the models of temporal point processes. A Cox process is defined via a stochastic intensity function. The stochastic process to generate the intensity function is usually chosen to be a Gaussian process (GP) and the model using a GP is called a Gaussian-process-modulated Poisson process (GP3) model. For the recurrent event data, GP3 models have been studied extensively. Among all approaches which try to solve the inference problem, the variational inference method provides a computationally efficient estimate of the intensity function and does not require a careful discretization of the underlying space.

In order to retain the scalability and computation efficiency of the variational inference approach and model the uncertainty of the intensity function when we

only observe panel count data, we present the first Bayesian inference framework for panel count data. We assume that all time-sequences in the data set share the same intensity function, which is generated by a GP3 model. The method of conducting computationally efficient variational inference is presented. We derive a tractable lower bound to alleviate the problem of the intractable evidence lower bound inherent in the variational inference framework. Our model, the Gaussian-process-modulated Poisson process for panel count data (GP4C), outperforms a non-Bayesian method in terms of the test likelihood and achieves comparable results in computation time.

For multiple time-sequences, it is often cumbersome to assume all time-sequences share the same intensity function since we may overlook the variety for different time-sequences. A key idea to model the heterogeneity is to cluster the data into groups while allowing the groups to remain linked to share the latent functions. Several models have been proposed on the basis of this simple idea, e.g., the convolution process, nonnegative matrix factorization (NMF), and latent Poisson process allocation (LPPA). These models employ latent factors to share statistical strengths and combine these functions to model the correlations within and among time-sequences. Among these models, LPPA is a powerful approach because it uses latent functions obtained from a GP, which is a flexible prior for a random function. However, a limitation of LPPA is that the number of latent functions needs to be set beforehand. If the chosen number is much larger than the actual number of latent functions required to explain the data, LPPA will still use all the latent functions and over-fit on the training data set.

To automatically infer the number of basis functions for multiple time-sequences, we present the Bayesian nonparametric Poisson process allocation (BaNPPA), a latent-function model for time-sequences. We model the intensity of each sequence as an infinite mixture of latent functions, each of which is obtained using a function drawn from a GP. We show that a technical challenge for the inference of such mixture models is the un-identifiability of the weights of the latent functions. We propose to cope with the issue by regulating the volume of each latent function within a variational inference algorithm. Our algorithm is computationally efficient and scales well to large data sets. We demonstrate the usefulness of our proposed model through experiments on both synthetic and real-world data sets.

In summary, we proposed two computationally efficient variational Bayesian inference algorithms for time-sequence modeling. In the first algorithm GP4C, we

quantified the average arrival rate for multiple time-sequences and provided the additional uncertainty, which helps illustrate the difficulty of the prediction. For the second algorithm BaNPPA, we automatically inferred the number of basis functions to model the variety for multiple time-sequences, which could provide insights into the understanding of social networks and human activities.