Fast Variational Inference for Dirichlet Process Mixture Models

Matteo Zanotto
Istituto Italiano di Tecnologia
Genova, Italy
matteo.zanotto@iit.it

Vittorio Murino
Istituto Italiano di Tecnologia
Genova, Italy
vittorio.murino@iit.it

Abstract

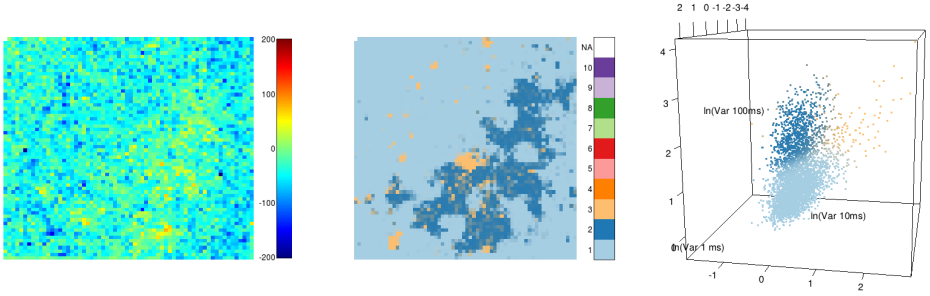
Nonparametric Bayesian Models currently suffer from a lack of efficient inference algorithms. This precludes the applicability of these methods when real-time analysis is needed. In this work we present a fast on-line variational inference algorithm for Dirichlet Process Mixture Models which takes advantage of the evolution of the data to speed up inference. The method has been applied to perform dynamic clustering on neuronal activity data in order to identify clusters having similar behaviour.

1 Introduction

The vast majority of applications of Nonparametric Bayesian Models require the use of MCMC methods in the inference stage. In order to reduce the computational burden of the inference process, a different approach based on variational inference has been proposed by Blei and Jordan [2]. Despite the considerable reduction of computation needed to get to acceptable accuracy levels, this approach is still unsuitable for real time applications. The method we propose builds onto the standard variational inference algorithm and takes advantage of the evolution of the considered process in order to distribute inference over time.

2 Fast Variational Inference for Dynamic Data

Many of the phenomena observed in nature have a smooth evolution over time. While several models have been proposed to explicitly model time evolution, none of them, to the best of our knowledge, exploits the dynamics of the data in order to reduce the amount of computation needed to perform inference.
This work reports a fast inference algorithm for Dirichlet Process Gaussian Mixture Models capable of dealing with streaming data coming from time-evolving processes. The mathematical formulation of the model is derived from the one proposed by Blei and Jordan [2] which has been revised integrating the formulas reported by Penny [3] in order to deal with a more general Gaussian-Wishart model. This extension allows to use Gaussian components without introducing any constraint on the shape of their covariance matrix which is fundamental for applicability to real-world problems.
The proposed algorithm can be seen a generalisation of the work by Blei and Jordan [2] which is obtained as a special case when the data do not change over time.
Whenever the observed process evolves in a smooth way (i.e. the process shows a certain extent of auto-correlation), this can be exploited to speed up inference by distributing computation over time. In practice, instead of iterating the variational updating formulas to convergence at every time step, one single update cycle is performed for each data-frame and the updated parameters are used to model the prior distribution for the following time step.



(a) (b) (c)

Figure 1: One sample of the raw signal in mV (1a), the correspondent clustering (1b) and the cluster representation in the considered 3D space (1c). Colours in 1b and 1c are obtained by interpolating the colour-bar elements according to the probability of each point to belong to each component.

3 Dynamic Clustering of Neuronal Signals

The proposed algorithm has been used to dynamically cluster neurons according to their electrical activity recorded with a high-density micro-electrode array sensing device [1]. Clustering neurons is of interest for neuroscientists in order to localise, in any given neuronal network, parts showing similar behaviour which suggests functional similarity of cells.
In our experiment, the signal of each electrode is represented, at each time-step, as a 3D vector whose components are the logarithm of the variance of the signal computed on time windows of 1, 10 and 100 ms, respectively. Variances have been used because they are well suited to describe the level of neuronal activity over a time period. The different lengths of the time windows allow to capture specific aspects of the signal evolution which are of interest from a biological point of view. Finally, logarithms have been taken since variances are lower-bounded by zero and hence badly approximable by Gaussians.
During the analysis the clustering evolves mirroring the evolution of the recorded signal and highlighting areas of the sensor which present a similar evolution. It is interesting to note how the obtained groups are locally compact despite the absence of any spatial constraint. The coherent evolution of the clusters over time, moreover, gives empirical evidence of the fact that the single iteration scheme used for the variational updates provides good results when inference is performed on a stream of data showing autocorrelation. An example of the obtained clusters is shown in Figure 1.
Due to the structure of the inference algorithm, the code is highly parallelisable. The current Matlab-based prototype takes about 0.03 seconds to process each data-frame (4,096 3D points) on a XEON E5620 2.4GHz processor and speed gains are expected from the C++ parallel implementation currently under development.

References

[1] Luca Berdondini, Kilian Imfeld, Alessandro Maccione, Mariateresa Tedesco, Neukom Simon, Milena Koudelka-Hep, and Sergio Martinoia. Active Pixel Sensor Array for High Spatio-Temporal Resolution Electrophysiological Recordings from Single Cell to Large Scale Neuronal Networks. *Lab on a Chip*, 9(18):2644–2651, 2009.

[2] David M. Blei and Michael I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[3] William D. Penny. Variational Bayes for d-dimensional Gaussian Mixture Models. Technical report, Wellcome Department of Cognitive Neurology - University College London, July 2001.