Online Bayesian Nonparametrics for Group Detection
Supplementary Material

Matteo Zanotto, Loris Bazzani, Marco Cristani, Vittorio Murino
Pattern Analysis & Computer Vision
Istituto Italiano di Tecnologia

1st August 2012

Abstract

This report complements the BMVC paper *Online Bayesian Nonparametrics for Group Detection* with additional information. In particular it presents the method designed to perform online clustering of streaming data with a Dirichlet Process Mixture Model. A sequential variational inference framework for time evolving data is presented and used to perform fast clustering with minimal memory requirements. The method has been tested on synthetic data and the produced results show that the method performs well while being extremely fast and memory efficient.

1 Introduction

Over the last decade, progressively more attention has been paid to the study of data coming form high-throughput processes. In several domains, moreover, it is important to process data streams in real-time in order to take decisions which modify the execution of a performed task. Examples of very different applications range from real-time control of experiments in biology, to real-time video analysis for monitoring and surveillance, to control of production processes, to monitoring of sensor networks. Most of these applications are characterised by the fact that the data streams of interest are extremely long and potentially endless which means that on-line analysis methods are more suitable than those based on batch elaboration in terms of both computational load and memory requirements.

One important data analysis task which has been extended to work on dynamic data is clustering. While the dynamics of the evolution of the data is very important in this task, in several situations modelling it explicitly a priori is either too complex or impossible.

In this report we present a method to perform dynamic clustering of contemporaneous observations (data frames) coming from data streams evolving in

parallel. The method performs clustering on-line and is based on sequential variational inference for Dirichlet Process Mixture Models.
The report is organised as follows: at first past relevant works will be briefly presented (Sec. 2), then the proposed method will be described (Sec. 3) and the obtained results will be discussed (Sec. 4). Finally conclusions will be drawn (Sec. 5).

2 Related Work

Clustering sequential data has been attracting attention in recent years and various techniques have been proposed to deal with the problem. As pointed out by several authors [12, 1], though, rather different approaches have been followed.
A first distinction must be made on the object of the clustering itself. A first class of methods [3] is aimed at clustering entire sequences. In this case the single data points within the sequence are not of interest in themselves and the whole set of points is considered as a single entity. The second class includes those methods for which the data points are the objects to be clustered. This class, generally known as dynamic clustering methods, is by far the most common and includes a series of different cases which differentiate themselves along two main directions: the way points are considered and whether only the past (on-line methods) or the entire sequence of observations (off-line methods) is needed to perform clustering. The distinction between on-line and off-line techniques is extremely important as it has a considerable impact on practical applications. While off-line methods take advantage of the whole sequences to find the best set of clustering configurations, on-line ones work on the basis of past and current information only. While this latter modality is more challenging and several complications arise, it is often the only applicable technique, especially if the output of the clustering is to be used to take decisions affecting the data-generating process itself (*e.g.* experiments or production processes to name a few).
Beringer and Hüllermeier [4] proposed and on-line method based on an extension of the K-means algorithm [9] to perform clustering of streaming data. Chakrabarti *et al.* [6] proposed a more flexible framework where the cost function minimised to define the clusters takes into account the smoothness in the evolution of clustering configurations over time. Their framework can work on-line and adapts easily to different types of clustering algorithms as shown in the paper. The main disadvantage of the method is the need for a user defined cost function which is often difficult to supply *a priori*. Both the mentioned methods can work on-line, but pay in terms of sophistication and richness of the model.
Conversely, more statistically sophisticated off-line methods have been proposed. Wang *et al.* [12] proposed the use of a Hidden semi-Markov Model to describe the dynamic evolution of the mixture models defining the data clusters, while Ahmed and Xing [1] presented the Temporal Dirichlet Process Mixture Model which extends the standard Dirichlet Process Mixture Models in order to con-

sider the evolution in time of the estimated components. While the model has several interesting properties, being an off-line method its application is limited to scenarios where entire data sequences are available at the time of analysis. The method presented in this work tries to give a contribution in bridging the highlighted gap, proposing an on-line method for dynamic clustering based on sophisticated probabilistic models like Dirichlet Process Mixture Models.

3 Sequential Variational Inference Framework

Approaching clustering as an inference procedure on a graphical model describing the data-generating process offers different advantages linked to the nature Bayesian inference. Among the different probabilistic models, Bayesian Nonparametrics have several properties which suit the application well. In particular, Dirichlet Process Mixture Models (DPMM) [2] represent mixture distributions with an unbounded number of components where the complexity of the model adapts to the observed data. This property is extremely important for dynamic clustering, as the number of clusters is, in general, not known *a priori* and can reasonably assumed to be changing over time. These models, in other words, can naturally manage situations like split/merge of clusters[1], creation of new clusters or suppression of obsolete ones which are expected to be common in dynamic evolution of data streams.

Despite their very appealing properties, Nonparametric Bayesian models are characterised by computationally-intensive inference procedures often based on Gibbs samplers. While Gibbs sampling can be an appropriate inference mechanism when execution time is not an issue, it is not applicable in situations where computation must be minimal in order to perform fast inference. As an alternative to Gibbs sampling variational inference can be used. Besides reducing the amount of computation needed, variational inference has interesting properties which make it more suitable than Markov Chain Monte Carlo (MCMC) approaches for the problem at hand. The most important of these properties is that variational inference maximises a lower bound to the true underlying distribution and so, after each iteration, the obtained parameters define a distribution which approximate the true one in a properly defined way. In contrast, MCMC provide a valid approximation only after convergence which generally takes many iterations and is anyway hard to assess.

This property of variational inference is particularly interesting when combined with some specific properties of sequences of streaming data. One of the peculiarities of data streams of practical interest is that their evolution is generally quite smooth i.e. could be seen as being generated by stochastic processes presenting some extent of (positive) autocorrelation. This translates in relatively small difference between two consecutive observations. In such a setting, it is reasonable to assume that, regardless the specific underlying process, what happened in the past could be used as a prior belief for what is going to happen

[1] In the paper we will use the expressions of cluster and mixture component to indicate the same concept, as already done by other authors [1].

next.

This observation, along with the properties of variational inference, suggests an efficient inference algorithm where, instead of iterating to convergence the variational updating formulas for each data frame, inference is distributed over time. In particular, one single iteration is performed for each update cycle and the obtained posterior is used as the prior for the next frame.

In practice, each cluster is seen as a component of a mixture model in the space where the observed points live. Variational Bayes is then used to update the parameters of the component distribution, defining the posterior parameters starting from the prior ones (posterior of the previous time step) and the current observations.

Performing single-iteration updates allows to speed up inference considerably with respect to MCMC-based methods which require to be run to convergence. Moreover, taking advantage of sequential inference this way, the dynamics of the clusters evolution is taken into account without explicitly modelling it. This is an interesting property in those situations where modelling the dynamics *a priori* is too complex or impossible altogether.

It is well known that results of variational approximation may be affected by the initialisation conditions as the Kullback-Leibler (KL) divergence can have local minima. While such minima are generally avoided by running repeated experiments with different initialisations, this process is time consuming and definitely not suitable when results have to be produced in real time. Empirical evidence gathered during experiments suggests that when running single update cycles the problem of local minima is not as significant as when variational updates are iterated to convergence. This is expected to happen because the dynamics of the observed data make the initialisation conditions less important after an initial transient. A formal analysis to thoroughly study this empirical observation will be subject of future work. While the presence of an initial transient is unavoidable, it is common to have calibration transients in many methods performing on-line analysis of time series data. Consequently this should not be seen as a strong limitation, especially considering that the method is thought to analyse long streams of data.

3.1 Mathematical Formulation

On the basis of the ideas introduced above, the problem of clustering a set of contemporaneous observations (data frame) has been formalised as an inference problem over a mixture model having infinitely many components. Each of the N_t points X_n observed at time t is generated, by one of the infinitely many components (k). It is important to underline that this approach deals naturally with a number of points possibly varying from data frame to data frame. Each component has its own parameters defining its distribution and has an associated probability mass depending on the parameters of the Dirichlet Process prior imposed on the components. For tractability, the Dirichlet Process prior [7] has been implemented as a stick-breaking construction [11].

Under this model, clustering is performed on the basis of the probability each

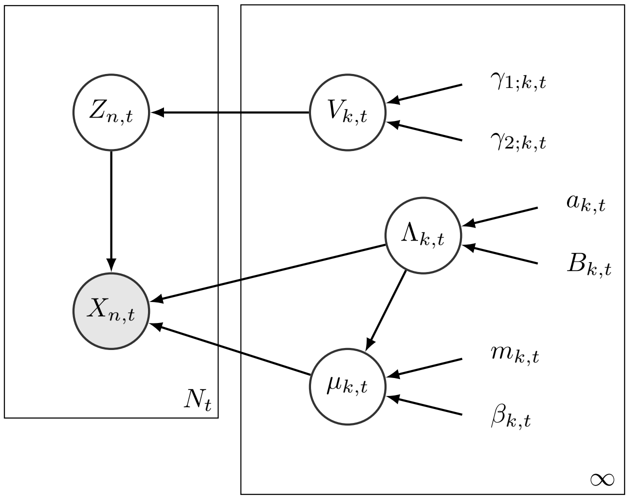

Figure 1: Graphical model representing the random variables and the hyper-parameters at time t. Hyper-parameters are included in the plate since they are not shared by the components. Refer to section 3.1 for details.

point has to belong to each component.
The graphical model associated to the described generative process is the one shown in Figure 1 where, at time step t, we have N_t points and

$$V_k|\gamma_{1;k},\gamma_{2;k} \sim \mathit{Beta}\left(\gamma_{1;k},\gamma_{2;k}\right) \tag{1}$$

$$Z_n|\{v_1,v_2,\ldots\} \sim \mathit{Discrete}\left(\pi\left(\mathbf{v}\right)\right) \tag{2}$$

$$\Lambda_k|\mathbf{B}_k,a_k \sim \mathit{Wishart}\left(\mathbf{B}_k,a_k\right) \tag{3}$$

$$\mu_k|m_k,\beta_k,\mathbf{\Lambda}_k \sim \mathit{Gaussian}\left(m_k,\left(\beta_k\mathbf{\Lambda}_k\right)^{-1}\right) \tag{4}$$

$$X_n|Z_n \sim \mathit{Gaussian}\left(\mu_{z_n},\mathbf{\Lambda}_{z_n}^{-1}\right) \tag{5}$$

where X_n represents the n^{th} data point, Z_n is an assignment variable relating each data point to the mixing components, V_k and the pair $(\mu_k,\mathbf{\Lambda}_k)$ represent the k^{th} mixture component in the stick-breaking construction [11] with $(\mu_k,\mathbf{\Lambda}_k)$ representing the location of the component in the parameter space and V_k defining the mixing proportions $\pi(\mathbf{v})$. The resulting stick-breaking construction G is

$$\pi_k(\mathbf{v}) = v_k\prod_{j=1}^{k-1}(1-v_j) \tag{6}$$

$$G = \sum_{k=1}^{\infty}\pi_k(\mathbf{v})\cdot\delta_{(\mu_k,\mathbf{\Lambda}_k)} \tag{7}$$

where $\delta_{(\mu_k,\mathbf{\Lambda}_k)}$ represent a Dirac delta function in the location defined by $(\mu_k, \mathbf{\Lambda}_k)$. Please note that pedix t has been dropped to keep notation uncluttered and will be introduced only when necessary.

It is important to underline that the proposed probabilistic formalisation employs conjugacy both on the observation side (the Gaussian-Wishart prior is conjugate to the Gaussian observation model) and on the components side (the Beta prior is conjugate to the Multinomial/Discrete distribution which defines the probability of each component). This has been done in order to speed up inference.

3.2 Variational Inference

Variational inference for Dirichlet Process Mixture Models has been originally proposed by Blei and Jordan [5]. Their work, though, focused on static data rather then on dynamic ones and variational updates were iterated to convergence. Moreover, the proposed algorithm was meant to deal with components having a fixed structure of the covariance matrix. In our work we overcame this latter limitation using a Gaussian-Wishart model for the mixture components, while single variational updates have been used to take advantage of the dynamics of the data, speeding up inference and incrementing the probability of avoiding local minima.

As proposed by Blei and Jordan [5], mean field variational inference can be formulated using a family of variational distributions over $\theta = \{\mathbf{v}, \mu, \mathbf{\Lambda}, \mathbf{z}\}$ based on a truncated stick breaking construction with truncation level K

$$q(\theta) = \prod_{k=1}^{K-1} q_{\gamma_k}(v_k) \prod_{k=1}^{K} q_{\tau_k}(\mu_k, \mathbf{\Lambda}_k) \prod_{n=1}^{N} q_{\phi_n}(z_n) \tag{8}$$

The introduction of this truncation level, if done appropriately, causes minimal approximation error while keeping the computation tractable [13]. In the formula above, n indexes the data points, k indexes the mixture components, $q_{\gamma_k} \sim Beta(\gamma_{1;k}, \gamma_{2;k})$, $q_{\tau_k}(\mu_k, \Lambda_k)$ follows a Gaussian-Wishart model parametrised by $\tau_k = \{m_k, \beta_k, \mathbf{B}_k, a_k\}$ such that $q_{\tau_k}(\mu_k, \Lambda_k) \sim \mathcal{N}\left(\mu_k | m_k, (\beta_k \mathbf{\Lambda}_k)^{-1}\right) \mathcal{W}(\mathbf{\Lambda}_k | \mathbf{B}_k, a_k)$ and $q_{\phi_k}(z_n) \sim Discrete(\phi_n)$. The product over the $q_{\gamma_{\cdot;k}}$ stops at component $K-1$ since the last component absorbs all the residual probability mass of the stick-breaking construction and hence $q_{\gamma_K}(v_K) = 1$ [5].

Variational Bayes inference takes the form of an Expectation-Maximisation algorithm and can be divided in an E-step and a M-step. The formulas of the two steps have been derived from the ones proposed by Blei and Jordan [5] and from those reported by Penny [10]. In the E-step the probability ϕ_n^k of each of the N points to belong to each of the K components is computed as:

$$\phi_n^k = \frac{exp\{S_n^k\}}{\sum_{j=1}^{K} exp\{S_n^j\}} \tag{9}$$

where

$$\begin{aligned} S_n^k = E_q[log(V_k)] + \sum_{i=1}^{k-1} E_q[log(1-V_k)] + \frac{1}{2} log\tilde{\Lambda}_k + \\ - \frac{1}{2}\left(x_n - m_k\right)' \bar{\Lambda}_t \left(x_n - m_k\right) - \frac{d}{2\beta_k} \end{aligned} \tag{10}$$

with

$$E_q[log(V_k)] = \Psi(\gamma_{1;k}) - \Psi(\gamma_{1;k} + \gamma_{2;k}) \tag{11}$$

$$\sum_{i=1}^{k-1} E_q[log(1-V_i)] = \Psi(\gamma_{2;k}) - \Psi(\gamma_{1;k} + \gamma_{2;k}) \tag{12}$$

$$\begin{aligned} log\tilde{\Lambda}_k &= \sum_{i=1}^{d} \Psi\left(\frac{a_k + 1 - i}{2}\right) + \\ &\quad - log\left|B_k\right| + d \cdot log\left(2\right) \end{aligned} \tag{13}$$

$$\bar{\Lambda}_k = a_k B_k^{-1} \tag{14}$$

with d being the dimensionality of the space and $\Psi(\cdot)$ being the digamma function.

Once all ϕ_n^k have been computed, the parameters of the distributions are updated in the M-step. After defining the following variables

$$\bar{N}_{k,t} = \sum_{n=1}^{N} \phi_n^{k,t} \tag{15}$$

$$\bar{\mu}_{k,t} = \frac{1}{\bar{N}_{k,t}} \sum_{n=1}^{N} \phi_n^{k,t} \cdot x_{n,t} \tag{16}$$

$$\bar{\Sigma}_{k,t} = \frac{1}{\bar{N}_{k,t}} \sum_{n=1}^{N} \phi_n^{k,t} \left(x_{n,t} - \bar{\mu}_{k,t}\right)\left(x_{n,t} - \bar{\mu}_{k,t}\right)' \tag{17}$$

the variational Bayes update formulas are used to update the parameters. The computed parameters define the posterior distributions at time-step t and will be used as prior for the following time-step from which the $t+1$ pedix is derived. In particular, the parameters $\gamma_{1;\cdot,t+1}$ and $\gamma_{2;\cdot,t+1}$ of the *Beta* distribution defining the sticks length and consequently the mixing proportions are updated as

$$\gamma_{1;k,t+1} = \gamma_{1;k,t} + \bar{N}_{k,t} \tag{18}$$

$$\gamma_{2;k,t+1} = \begin{cases} \gamma_{2;k,t} + \sum_{j=k+1}^{K} \bar{N}_{j,t} & \text{if } k < K \\ \alpha & \text{if } k = K \end{cases} \tag{19}$$

with α being the scaling constant of the Dirichlet Process prior at time-step 0 [5]. The parameters of the distribution of the mean of each component of the mixture are updated as

$$m_{k,t+1} = \frac{\bar{N}_{k,t} \cdot \bar{\mu}_{k,t} + \beta_{k,t} \cdot m_{k,t}}{\bar{N}_{k,t} + \beta_{k,t}} \tag{20}$$

$$\beta_{k,t+1} = \bar{N}_{k,t} + \beta_{k,t} \tag{21}$$

Finally the parameters of the distribution of the precision matrix of each component are updated according to

$$a_{k,t+1} = \bar{N}_{k,t} + a_{k,t} \tag{22}$$

$$\begin{aligned} B_{k,t+1} = {} & \bar{N}_{k,t} \cdot \bar{\Sigma}_{k,t} + \\ & + \frac{\bar{N}_{k,t} \cdot \beta_{k,t} \left(\bar{\mu}_{k,t} - m_{k,t}\right) \left(\bar{\mu}_{k,t} - m_{k,t}\right)'}{\bar{N}_{k,t} + \beta_{k,t}} + \\ & + B_{k,t} \end{aligned} \tag{23}$$

After the parameter initialisation, the variational E-M procedure can be applied to the incoming data streams to perform inference over time.

After the conclusion of the M-step, the components which are not maximally responsible for at least one point (i.e. components $\bar{k}$ such that, $\nexists$ point n such that $\bar{k} = \mathit{arg\,max}_j\, \phi_n^j$) are considered unassigned and re-initialised to have mean in areas of the space badly modelled by the current mixture. To conclude the variational update, the components are sorted by decreasing number of assigned points.

3.3 Properties of the Model

The probabilistic model proposed presents several desirable properties which make it extremely well suited for online clustering of streaming data.

The first and most prominent property is that the model processes one single data frame at each time-step and uses the observations, along with the prior parameters, to derive the posterior mixture model. Integrating observations over time through sequentially iterating this process, has two main advantages: on the one hand it imposes a minimal computation burden, on the other hand it has limited memory requirements which remain constant regardless the length of the data stream. These two advantages make the proposed method well suited for situations in which either speed is a major concern or long and potentially endless data streams are to be processed.

Finally, given its algorithmic structure the inference process is highly parallelisable. This fact is particularly interesting considering that implementations on GPUs could allow to perform real-time clustering of high-throughput data streams.

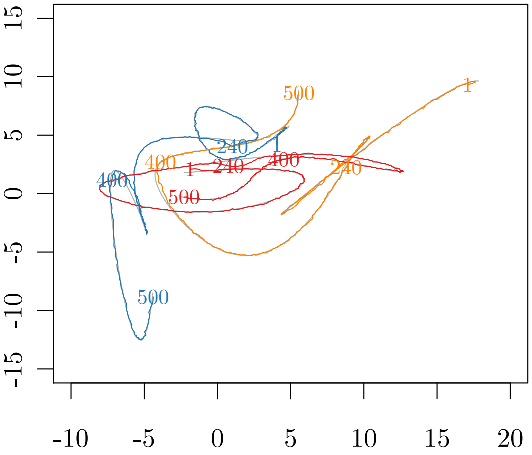

Figure 2: Real (grey) and estimated (colour) trajectories of the cluster means in one specific run. Labels mark the position of the means at specific time steps. Trajectories are nearly overlapping and differences are better appreciated zooming on screen.

4 Experimental Results

The proposed clustering technique has been tested on a synthetic dataset to assess the properties of the model and verify its ability to retrieve the real underlying clusters.

In order to assess the ability of the proposed method to consistently identify clusters evolving over time a synthetic dataset has been generated. The parallel data streams have been simulated with a set of samples drawn from Gaussian distributions evolving over time and the method was evaluated on the basis of how well it could recover the underlying generating distributions.

A set of three 2D Gaussian distributions moving in space has been considered. For each of them, the trajectory of the mean through space has been simulated by generating a set of points from an ARMA(1,1) model (Autoregressive Moving Average model of order 1) [8] with high AR and MA coefficients (both set to 0.95) and then performing a spline interpolation between them. This way the generated trajectories were autocorrelated and smooth.

Cluster trajectories have been generated to simulate most of the challenges

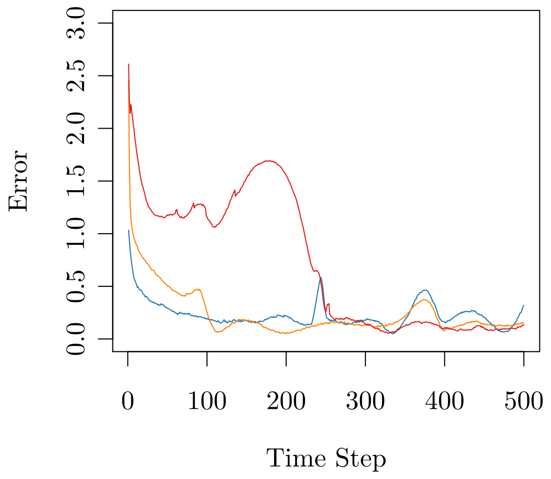

Figure 3: Error between estimated and real mean for each cluster (average over 50 experiments). The estimation error drops after an initial transient showing that the method correctly infers the position of the clusters.

which characterise real situations. In particular, trajectories are very complex and occasionally clusters have considerable overlap resembling a single distribution.

At each time-step t the algorithm was presented with a collection of data points (data frame) sampled from the three Gaussians whose mean was the t^{th} point in the related trajectory. The covariance matrix was kept constant for easier interpretation of the results, but everything extends to changing covariances without any modification.

In order to obtain meaningful statistics 50 experiments have been run using the same cluster trajectories but drawing each time different samples from the Gaussian distributions. In all the experiments the truncation level of the stick-breaking construction was fixed to 20.

Figure 2 shows the real and estimated trajectories of the means, while Figure 3 reports the error between the two. As previously mentioned, in the initial few frames (roughly 100 in this case) the approximation is coarser. This is due to the fact that the three generating components have not been correctly identified because of the limited information cumulated during sequential inference and for the effect of the initialisation of the parameters. Once enough evidence is absorbed in the model parameters, the generating clusters are correctly outlined and the error drops to low levels.

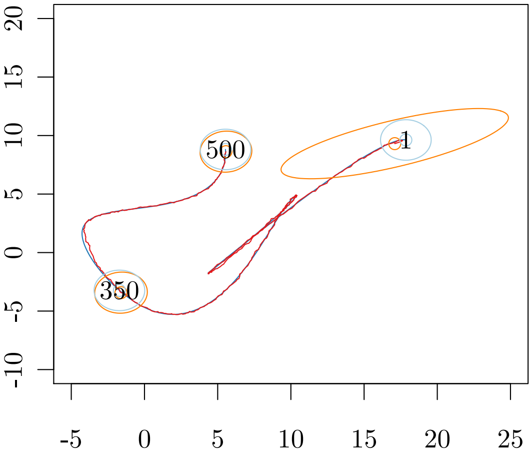

Figure 4: Close up on a single trajectory showing real (blue) and estimated trajectory (red). Real (light blue) and estimated (orange) covariances are shown at three time steps (1, 350, 500) as indicated by the black labels.

The higher level of error for the components associated to the orange and red trajectory is explained by a difficult situation deliberately induced to test the strength of the method. When two components get very near and the generated clusters have substantial overlap, it is difficult to decide how to assign the cluster labels once the two trajectories get separated again. Over the 50 experiments, occasional switches of labels have been observed in early time-steps between clusters associated to orange and red trajectories which motivates the higher level of error than that observed for the blue one.

The mentioned overlap between clusters is the source of error behind the peaks observed after time-step 200 and around 400. When clusters overlap, in fact, the assignment of points to mixture components is impossible without additional information as many points could have been equally likely generated by either mixture component. While there is no trivial solution to this problem, it is important to notice that the proposed model is only marginally affected by this challenging situation since only a small increase in the error is observed and the components are kept separate and not fused into a single bigger cluster. A notable exception is one of the observed trials when the cluster associated to the red trajectory has been moved way off up to frame 200. Apart form this case,

Table 1: Root Mean Square Error between real and estimated trajectories for complete sequence (RMSE-2) and after the initial 100 frames (RMSE-1). Tracks are identified by their colour in the graph. Reported data are averages over 50 experiments.

	Blue	Orange	Red
RMSE-1	0.1969	0.1448	0.7388
RMSE-2	0.2585	0.3349	0.9592

which explains the high level of error for the red trajectory in the initial frames, the proposed method is extremely effective in dealing with the complex task faced. The result is even more interesting when considering that the method processes single data frames and takes advantage of past data only through the prior.

Figure 4, finally, reports a close-up on a single trajectory, showing how the estimated covariance matrix evolves over time.

In order to evaluate quantitatively the obtained results, Table 1 reports the Root Mean Square Error computed for the three trajectories. Both the values ignoring the initial 100-frame transient (RMSE-1) and those for the complete trajectories (RMSE-2) are satisfactorily low when compared to the variability range of x and y coordinates (see Figure 2) and witness the ability of the method to retrieve the original clusters.

From a more general point of view, it is important to underline that the method was able to identify the correct number of clusters after the initial transient and correctly prevented them to be merged in complex situations on the basis of the past evidence summarised in the prior distribution.

5 Conclusions

An online method for dynamic clustering of data streams has been proposed. The method builds on Dirichlet Process Mixture Models, exploiting their ability to model mixture distributions having an unbounded number of clusters, and on a sequential variational inference framework able to take advantage of the dynamics of the clusters without explicitly modelling it.

The method has been tested on challenging data created for the purpose. The proposed method performed well, showing its ability to infer the correct parameters of the evolving clusters in the synthetic dataset.

The results obtained in the experiments, the theoretical properties of the model and the fact that inference is highly parallelisable, show that the method could be promising for real-time clustering of high-throughput data streams.

References

[1] Amr Ahmed and Eric Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process : with applications to evolutionary clustering. In *Proceedings of The Eighth SIAM International Conference on Data Mining (SDM2008)*, pages 219–229, 2008.

[2] Charles E. Antoniak. Mixtures of Dirichlet Processes with Applicatioins to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[3] Ziv Bar-Joseph, Georg Gerber, David K. Gifford, Tommi S. Jaakkola, and Itamar Simon. A new approach to analyzing gene expression time series data. In *Proceedings of the International Conference on Computational Biology*, pages 39–48, 2002.

[4] Jürgen Beringer and Eyke Hüllermeier. Online clustering of parallel data streams. *Data & Knowledge Engineering*, 58(2):180–204, August 2006.

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

[6] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[7] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, March 1973.

[8] James D. Hamilton. *Time Series Anlaysis*. Princeton University Press, 1994.

[9] John A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

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

[11] Jayaram Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

[12] Yi Wang, Shi-Xia Liu, Jianhua Feng, and Lizhu Zhou. Mining naturally smooth evolution of clusters from dynamic data. In *Proceedings of The Seventh SIAM International Conference on Data Mining (SDM2007)*, pages 125–134, 2007.

[13] Oliver Zobay. Mean Field Inference for the Dirichlet Process Mixture Model. *Electronic Journal of Statistics*, 3:507–545, 2009.