Scalable Inference in Dynamic Mixture Models

Patrick Jähnichen, Florian Wenzel, Marius Kloft
Machine Learning Group
Department of Computer Science
Humboldt-Universität zu Berlin, Germany
{patrick.jaehnichen, wenzelfl, kloft}@hu-berlin.de

Abstract

Previous work on inference for dynamic mixture models has so far been directed to models that follow a simple Brownian motion diffusion over time and pursued a batch inference approach. We generalize the underlying dynamics model to follow a Gaussian process, introducing a novel class of dynamic priors for mixture models. Further, we propose a stochastic variational inference scheme and compare our approach to previous solutions in terms of runtime complexity and test error.

1 Introduction

Despite their extraordinary capabilities to describe complex behavior in data, dynamic mixture models are not as heavily used as their static counterparts. Introducing dynamics to mixture models allows us to keep track of mixture components that are subject to a drift. Examples include the analysis of stock market data or time-stamped document collections (i.e. dynamic topic models) and weather forecasting, among others. In our approach, the underlying dynamics are modelled via Gaussian processes (GPs), opening up for a wide range of dynamic priors in mixture models and models of mixed membership. These include Brownian motion, the Ornstein-Uhlenbeck process (being the continuous AR(1) model) and periodic process priors. Further, we develop scalable inference methods for this new model class.

2 Dynamic Mixture Model

We study a novel modelling class introducing new kinds of dynamic priors for mixture models on time series. The model under study is a mixture model of L D-dimensional Gaussian distributions whose time-dependent dynamics are governed by a GP as described by the generative process:

1. for all $l = 1, \dots, L$ draw $\beta_l \sim \mathcal{GP}(0, K)$
2. for all $t = 1, \dots, T$ draw $\theta_t \sim \text{Dir}_L(\alpha)$
3. for all $n = 1, \dots, N$
   (a) draw a component: $z_n \sim \text{Mult}(\theta_{t_n})$
   (b) draw data $x_n \sim \mathcal{N}(\beta_{z_n, t_n}, \sigma_X^2 \mathbf{I})$,

where β_l are mixture components (each of which is a time-series over T steps), as given by a zero-mean GP prior with kernel function $k(\cdot, \cdot)$ and associated covariance matrix K. θ_t denotes the prior over mixing proportions for each data point at time t, σ_X^2 is a variance parameter and t_n is the observed

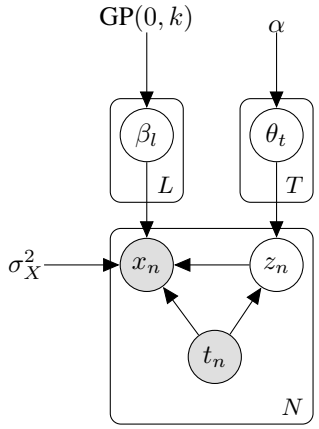


Figure 1: A simple GP dynamic mixture model.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

time-stamp associated with observation x_n. We give a graphical representation of the model in Fig. 1. We emphasize that with the flexibility to easily employ different kernel functions, we are able to capture a wide range of dynamic behavior of the data (e.g. linear and non-linear drifts, periodicity and jumps). Using e.g. a Wiener kernel function[1] implies that the mixture component means underlie a Brownian motion diffusion over time leading to a hierarchical model in the spirit of [1]. [2] describe a class of dynamic mixture models based on linear dynamics, while we stick to the idea of random walk state-space models and generalize those using GPs.

3 Related work

Our main focus lies on efficient inference in the described model. Bayesian inference in dynamic mixture models so far mainly relies on either MCMC sampling techniques (e.g. [3, 2]) or variational methods (using a variational interpretation of the Kalman-Filter (VKF) as in [4, 1]). All of these are batch inference algorithms and we improve upon them in terms of computation time by following a stochastic gradient descent approach. Recently, [5] introduced a stochastic MCMC sampling approach but is restricted to the enclosed case of models based on Brownian motion diffusion.

4 Inference

We develop both a batch and stochastic variational inference scheme for our model. While [1]'s VKF approach is custom-tailored to the Brownian motion diffusion case, our GP based approach does not suffer from this limitation.

4.1 Batch model

From the above definitions we construct a lower bound on the evidence (ELBO) by following standard variational mean field theory [6]. As our model is fully conjugate, we are able to derive full conditionals for all variables involved. We introduce variational distributions on each θ_t, $q(\theta_t|\lambda_t) = \text{Dir}_L(\lambda_t)$ and each z_n, $q(z_n|\phi_n) = \text{Mult}(\phi_n)$. Further, we place a T-dimensional variational distribution on the mixture components time series, $q(\beta_l) = \mathcal{N}_T(m_l, S_l)$[2]. Doing so yields a batch variational inference algorithm with the following ELBO objective and parameter updates:

$$\mathcal{L} = \mathbb{E}_q[\log p(\theta, \beta, z, x)] - \mathbb{E}_q[\log q(\theta, \beta, z)]$$

$$\phi_{nl} \propto \exp\left\{\psi(\lambda_{t_n,l}) - \psi\left(\sum_{l'} \lambda_{t_n,l'}\right) - \frac{1}{2\sigma_X^2}\left((x_n - m_l^{t_n})^T(x_n - m_l^{t_n}) + D(S_l)_{t_n,t_n}\right)\right\} \quad (1)$$

$$\lambda_{tl} = \alpha + \sum_n \mathbb{1}_{[t=t_n]}\phi_{nl} \quad (2)$$

$$m_l = \left(K_{TT}^{-1} + \frac{1}{2\sigma_X^2}\Phi_l\right)^{-1} \frac{1}{2\sigma_X^2}\Xi_l, \quad S_l = \left(K_{TT}^{-1} + \Phi_l\right)^{-1} \quad (3)$$

where K_{TT} is the covariance function evaluated on all observed time stamps, $\mathbb{1}_{[\cdot]}$ is the indicator function. Φ_l is a diagonal $T \times T$-matrix with $(\Phi_l)_{t,t} = \sum_n \mathbb{1}_{[t=t_n]}\phi_{nl}$ and Ξ_l is a $T \times D$-matrix with the t-th row being $\sum_n \mathbb{1}_{[t=t_n]}\phi_{nl}x_n^T$. Note that we are assuming independence in the individual dimensions of x_n and so are able to handle all dimensions simultaneously by using matrix algebra where appropriate.

4.2 Scalable model

To scale to considerably larger amounts of data we follow the ideas in [7] and consider a set of inducing variables, $\hat{\beta}$, which contain function values[3] at inducing locations $\mathbf{z} = \{z_i\}_{i=1}^I$ with $I < T$.

[1] Wiener kernel function: $k(t_i, t_j) = \min(t_i, t_j)$

[2] Note that we have to handle each dimension separately in this case and assume independence between dimensions.

[3] We utilize the function view on GPs by assuming that the mixture component means are functions of time.

We will use these to construct a lower-rank approximation to the underlying GP. To this end, we employ a prior on $\hat{\beta}$, $p(\hat{\beta}) = \mathcal{N}(0, K_{II})$ and, using standard GP results, we obtain

$$p(\beta^{(l)}|\hat{\beta}^{(l)}) = \mathcal{N}(K_{TI}K_{II}^{-1}\hat{\beta}^{(l)}, \tilde{K}) \tag{4}$$

with K_{II} the matrix resulting from evaluating the covariance function between all I inducing points. Here K_{TI} is the cross-covariance between the data points and the inducing points and $\tilde{K}$ is given by $\tilde{K} = K_{TT} - K_{TI}K_{II}^{-1}K_{IT}$. Applying Jensen's inequality on $p(x_n|z_n, t_n, \beta)$ and computing expectations under Eq. 4, we obtain

$$\begin{aligned}
\log p(x_n|z_n = l, t_n, \hat{\beta}) &= \log \mathbb{E}_{p(\beta|\hat{\beta})}\left[p(x_n|z_n, t_n, \beta)\right] \\
&\geq \mathbb{E}_{p(\beta|\hat{\beta})}\left[\log p(x_n|z_n, t_n, \beta)\right] \\
&= \log \mathcal{N}(k_{t_n,I}K_{II}^{-1}\hat{\beta}^{(z_n)}, \sigma_X^2) - \frac{1}{2\sigma_X^2}\tilde{k}_{t_n,t_n} \triangleq \mathcal{L}_1
\end{aligned} \tag{5}$$

where $k_{t_n,I}$ is the t_n-th row of K_{TI}. Thus, after incorporating the remaining parts of the model, our final objective is given by

$$\mathcal{L}_2 = \mathbb{E}_q\left[\sum_t(\log p(\theta_t|\alpha) - \log q(\theta|\lambda)) + \sum_n \log p(z_n|\theta_{t_n}) - \log q(z_n|\phi_n) + \mathcal{L}_1 + \log p(\hat{\beta}) - \log q(\hat{\beta})\right].$$

Here, $q(\hat{\beta}) = \prod_i \mathcal{N}(\hat{\beta}^{(i)}|m_i, S_i)$ is the variational distribution on $\hat{\beta}$, i.e. the variational parameters m and S now govern the approximating GP defined by the inducing function values $\hat{\beta}$. We now proceed by randomly selecting a subset $\mathcal{S}$ of the data and then updating local variables for this mini-batch. We use these local updates for constructing a noisy gradient on the global variables leading to a stochastic gradient descent scheme.

Updating local variables The parameter updates for local variables are similar to Eq. 1, differing only in the likelihood term which is now an expectation of $\mathcal{L}_1$ (Eq. 5) under the variational distribution,

$$\phi_{nl} \propto \exp\left\{\psi(\lambda_{t_n,l}) - \psi\left(\sum_{l'}\lambda_{t_n,l'}\right) - \frac{1}{2\sigma_X^2}\left((x_n - \mu_{l,t_n})^T(x_n - \mu_{l,t_n}) + \text{tr}(S_l\Lambda_{t_n}) + \tilde{k}_{t_n,t_n}\right)\right\}$$

where μ_{l,t_n} is given by $k_{t_n,I}K_{II}^{-1}m_i$, $\Lambda_{t_n} = K_{II}^{-1}k_{t_n,I}^T k_{t_n,I}K_{II}^{-1}$ and $\text{tr}(\cdot)$ is the trace operator.

Updating global variables The global variables in our model are θ_t and $\hat{\beta}_l$. For each θ_t we can make use of the fact that the natural gradient is identical to the standard coordinate ascent updates (Eq. 2). The update in step s is then

$$\lambda_{t,l}^{(s+1)} = \lambda_{t,l}^{(s)} + \rho_s\left(\alpha + \frac{N}{|\mathcal{S}|}\sum_{n=1}^{N}\mathbb{1}_{[t=t_n]}\phi_{n,l}\right)$$

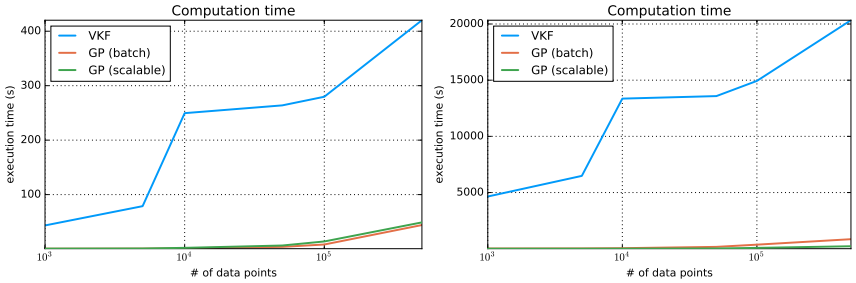
where ρ_s is a decreasing learning rate.Updating the parameters m_l and S_l follows along these lines. We can use the fact, that the gradient in expectation parameters automatically yields the natural gradient in canonical parameters in any exponential family distribution [8]. Thus we reparameterize $q(\hat{\beta}^{(l)}|m_l, S_l)$ by $\eta_l^{(1)} = S_l^{-1}m_l$ and $\eta_l^{(2)} = -\frac{1}{2}S_l^{-1}$ and construct the gradients of $\mathcal{L}_2$) in terms of expectation parameters to obtain the natural gradient. These are given by

$$\begin{aligned}
\frac{\partial\mathcal{L}_2}{\partial m_l} &= \frac{N}{|\mathcal{S}|}\sum_n \mathbb{1}_{[t=t_n]}\phi_{nl}K_{II}^{-1}k_{I,t_n}x_n - \Lambda m_l \\
\frac{\partial\mathcal{L}_2}{\partial S_l} &= \frac{1}{2}S_l^{-1} - \frac{1}{2}\Lambda
\end{aligned}$$

where $\Lambda = K_{II}^{-1} + \frac{1}{\sigma_X^2}\frac{N}{|\mathcal{S}|}\sum_n \phi_{nl}K_{II}^{-1}k_{I,t_n}k_{I,t_n}^T K_{II}^{-1}$. Updating the canonical parameters

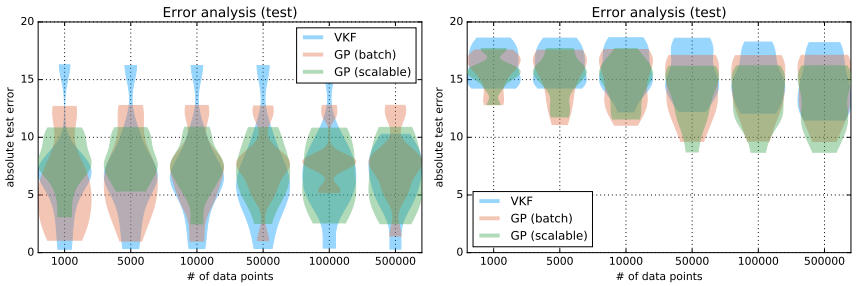
$$\begin{aligned}
\eta_l^{(1)(s+1)} &= \eta_l^{(1)(s)} + \rho_s\frac{\partial\mathcal{L}_2}{\partial m_l} \\
\eta_l^{(2)(s+1)} &= \eta_l^{(2)(s)} + \rho_s\frac{\partial\mathcal{L}_2}{\partial S_l}
\end{aligned}$$

completes the inference procedure.



(a) Simple model. (b) Complex model.

Figure 2: Runtime statistics.



(a) Simple model. (b) Complex model.

Figure 3: Test error statistics.

5 Experiments

We evaluate our approach on synthetic data generated according to our generative model with Wiener kernel function and use two different settings ($T = 10, D = 5, L = 5$ and $T = 100, D = 50, L = 25$). For both, we measure statistics for a growing number of observations N and collect run-time (Figures 2a and 2b) and test error statistics (Figures 3a and 3b). Here, VKF denotes the variational Kalman filter approach as introduced in [4], GP (batch) and GP (scalable) are the two inference schemes described above. For the latter two, we use the Wiener covariance function to be comparable to the VKF approach. As would be expected, the VKF and the batch GP algorithm perform with similar performance, although the latter is clearly faster in terms of computation time. This can be explained by the need of numerical optimization in the VKF, while the batch GP uses a direct coordinate ascent update. For the simpler problem, our scalable GP algorithm performs slightly less accurate in terms of predictive quality. As it uses a lower rank approximation to the resulting covariance matrix of the full batch GP approach this is again expected behavior. With increasing model complexity, the batch GP approach is still much faster than the VKF, however, the stochastic variational inference approach benefits from utilizing a lower-rank approximation and its property to reach an optimum after having processed much less data points than needed by a batch algorithm.

6 Discussion and Future Work

We explore new kinds of dynamic priors for Bayesian dynamic mixture models and thereby study a new modeling class. This opens up for utilizing well known dynamic priors in context of mixture models (e.g. the OU process). Further, we propose a stochastic variational inference scheme and find that it performs superior to the VKF in terms of computation time making it applicable to huge data sets. Our aim is to apply our findings to more complex models of mixed membership, especially topic models [9], leading to a truly scalable inference scheme for dynamic topic models *and* to the possibility of incorporating a broader range of prior assumptions on the type of diffusion for topics. Additionally, this formulation also allows to place priors on any hyperparameters as well, leading to a model that can capture known phenomena in time series analysis such as jumps, heteroscedasticity and stochastic volatility.

Acknowledgments

We thank Stephan Mandt for fruitful discussions. This work was partly funded by the German Research Foundation (DFG) award KL 2698/2-1 and the German Ministry of Education and Research (BMBF) within the ID:SEM research program, project PREDICT (031L0023A).

References

[1] Chong Wang, David M Blei, and David Heckerman. Continuous Time Dynamic Topic Models. In *Conference on Uncertainty in Artifical Intelligence*, 2008.

[2] C Glynn, S T Tokdar, D L Banks, and B Howard. Bayesian Analysis of Dynamic Linear Topic Models. *arXiv.org*, 2015.

[3] Richard Gerlach, Chris Carter, and Robert Kohn. Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, 95(451):819–828, September 2000.

[4] David M Blei and John D Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[5] A Bhadury, J Chen, J Zhu, and Shixia Liu. Scaling up Dynamic Topic Models. In *Proceedings of the 25th International Conference on the World Wide Web*, 2016.

[6] Martin J Wainwright and Michael I Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2007.

[7] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian Processes for Big Data. In *Conference on Uncertainty in Artifical Intelligence*, 2013.

[8] James Hensman, Magnus Rattray, and Neil D Lawrence. Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems*, 2012.

[9] David M Blei and John D Lafferty. Topic models. In Ashok N Srivastava and Mehran Sahami, editors, *Text Mining: Classification, Clustering, and Applications*, page 71. CRC Press, 2009.