Scalable Inference in Dynamic Mixture Models

Patrick Jähnichen, Florian Wenzel and 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 and test error.

Introduction

- Dynamic mixture models are not as heavily used as their static counterparts in spite of their ability to capture higher complexity in the data
- Dynamics in mixture models allow us to keep track of mixture components that are subject to a drift
  - Stock market data analysis
  - Time-stamped document collections (i.e. dynamic topic models)
  - Weather forecasting
- Our approach: model the underlying dynamics via Gaussian processes (GPs)
- Opens up for a wide range of dynamic priors in mixture models and models of mixed membership
  - Includes "classical" case of Brownian motion
  - Ornstein-Uhlenbeck process (the continuous AR(1) model)
  - Periodic process priors
- We develop a scalable inference method for this new model class

Standard Dynamic Mixture Models

Generative process

1. for all $l = 1, \dots, L$
   (a) draw $\beta_{l,0} \sim \mathcal{N}(\mu_0, \sigma_0^2)$
   (b) for all $t = 1, \dots, T$ draw $\beta_{l,t} \sim \mathcal{N}(\beta_{l,t-1}, \nu^2 \Delta_{t,t-1})$
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 $x_n \sim \mathcal{N}(\beta_{z_n,t_n}, \sigma_X^2 \mathbf{I})$,

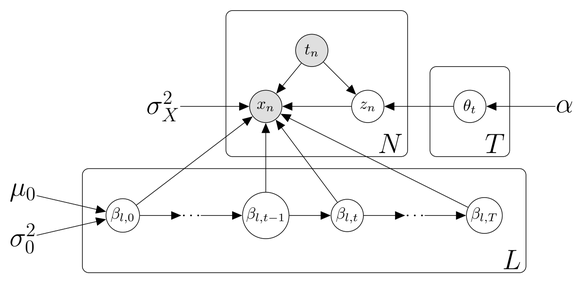

Fig. 1: A simple dynamic mixture model.

- Mixture model of L D-dimensional jointly Gaussian time series of length T in the spirit of [4]
- Time series dynamics governed by a first order Markov chain
- $\beta_{l,t}$ is mixture components l at time t
- θ_t denotes the prior over mixing proportions for each data point at time t
- t_n is the observed time-stamp associated with observation x_n
- σ_X^2 is the data variance parameter
- Identical to assuming Brownian motion diffusion through time on mixture components with variance parameter ν^2
- State-of-the-art variational inference method is Variational Kalman Filtering (VKF) as introduced in [1]

GP Dynamic Mixture Models

- Time series dynamics now governed by a general Gaussian process
- β_ls are given by a T-dimensional zero-mean GP prior with kernel function $k(\cdot, \cdot)$ and associated covariance matrix K
- Gives flexibility to easily employ different kernel functions and capture a wide range of dynamic behavior of the data
- Using the Wiener kernel function ($k(t_i, t_j) = \min(t_i, t_j)$) is identical to teh model above

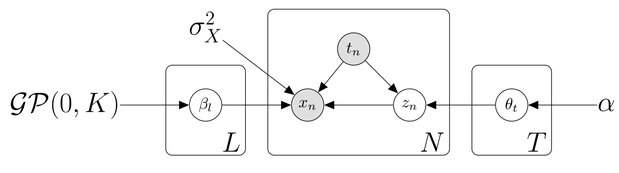

Fig. 2: The GP dynamic mixture model.

Inference

Batch algorithm

Variational family

- Introduce variational distributions on hidden variables
  - for all $t = 1, \dots, T$ set $q(\theta_t|\lambda_t) = \text{Dir}_L(\lambda_t)$
  - for all $n = 1, \dots, N$ set $q(z_n|\phi_n) = \text{Mult}(\phi_n)$
  - for all $l = 1, \dots, L$ set $q(\beta_l) = \mathcal{N}_T(m_l, S_l)$
- Variational distribution factorizes completely, save $\beta_{l,1:T}$

Parameter updates

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

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

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

- K_{TT} is the covariance function evaluated on all observed time stamps
- $\mathbb{1}_{[\cdot]}$ is the indicator function
- Φ_l and Ξ_l are the sufficient statistics to the variational distribution on β_l
  - Φ_l is a diagonal $T \times T$-matrix with $(\Phi_l)_{t,t} = \sum_n \mathbb{1}_{[t=t_n]}\phi_{nl}$
  - Ξ_l is a $T \times D$-matrix with the t-th row being $\sum_n \mathbb{1}_{[t=t_n]}\phi_{nl}x_n^T$

Scalable algorithm

Low-rank inducing point model

- Utilize stochastic variational inference on a lower-rank model using inducing points [2]
- Consider a set of inducing variables, $\hat{\beta}$ at inducing locations $\mathbf{z} = \{z_i\}_{i=1}^I$ with $I < T$
- Let $\hat{\beta} \sim \mathcal{GP}(0, K_{II})$ be a lower-rank GP prior
- Approximate full-rank GP using $\hat{\beta}$

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

- K_{II} is the inducing point covariance matrix
- K_{TI} is the cross-covariance between data points and inducing points
- $\tilde{K} = K_{TT} - K_{TI}K_{II}^{-1}K_{IT}$
- Introduce variational distribution on $\hat{\beta}$, $q(\hat{\beta}) = \prod_i \mathcal{N}(\hat{\beta}^{(i)}|m_i, S_i)$
- Apply Jensen's inequality on data likelihood $p(x_n|z_n, t_n, \beta)$

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

- $k_{t_n,I}$ is the t_n-th row of K_{TI}
- Final objective is now a lower bound to the "traditional" ELBO

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

- Proceed with stochastic variational inference (SVI) [3] scheme by randomly selecting minibatches $\mathcal{S}$ and optimizing $\mathcal{L}_2$ using noisy gradients

Parameter updates

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

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

- $\Lambda_t = K_{II}^{-1}k_{I,t}k_{I,t}^T K_{II}^{-1}$
- $\Lambda = K_{II}^{-1} + \frac{1}{\sigma_X^2}\frac{N}{|\mathcal{S}|}\sum_t \sum_n \mathbb{1}_{[t=t_n]}\phi_{nl}\Lambda_t$
- Use exponential family property $\nabla_\theta \mathcal{L}(\theta) = \tilde{\nabla}_\eta \mathcal{L}(\eta)$ and update canonical parameters η

$$\frac{\tilde{\partial}\mathcal{L}_2}{\partial \eta_l^{(1)}} = \frac{\partial \mathcal{L}_2}{\partial m_l} = \frac{N}{|\mathcal{S}|}\sum_n \mathbb{1}_{[t=t_n]}\phi_{nl}K_{II}^{-1}k_{I,t_n}x_n - \Lambda m_l$$

$$\frac{\tilde{\partial}\mathcal{L}_2}{\partial \eta_l^{(2)}} = \frac{\partial \mathcal{L}_2}{\partial S_l} = \frac{1}{2}S_l^{-1} - \frac{1}{2}\Lambda$$

$$\eta_l^{(1)(s+1)} = \eta_l^{(1)(s)} + \rho_s\frac{\tilde{\partial}\mathcal{L}_2}{\partial \eta_l^{(1)}}$$

$$\eta_l^{(2)(s+1)} = \eta_l^{(2)(s)} + \rho_s\frac{\tilde{\partial}\mathcal{L}_2}{\partial \eta_l^{(2)}}$$

Results

- Evaluation on two artificial data sets
  - Simple model: $T = 10, D = 5, L = 5$
  - Complex model: $T = 100, D = 50, L = 25$
  - $N \in \{1000, 5000, 10.000, 50.000, 100.000, 500.000\}$
- Number of inducing point for SVI approach is fixed to $I = 10$
- VKF and batch GP algorithm perform similar, latter is clearly faster
- Scalable GP algorithm slightly less accurate in predictive quality for simpler problem
- For increasing model complexity, batch GP approach still much faster than VKF, but SVI approach benefits from lower-rank approximation and reaching optimum after seeing less data points

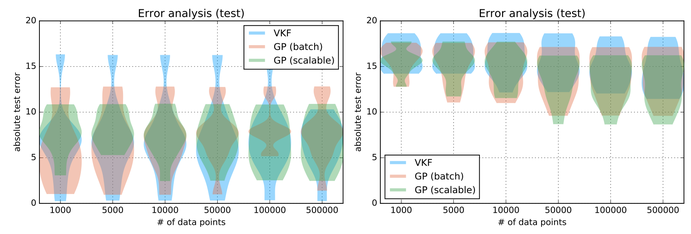

Fig. 3: Test error statistics. Left:

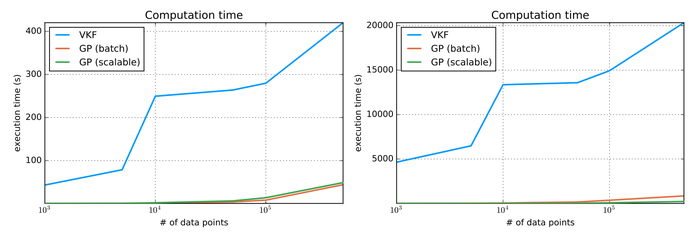

Fig. 4: Computation time statistics.

Contribution

- Explore new kinds of dynamic priors for Bayesian dynamic mixture models and thereby study a new modeling class
- Opens up for utilizing well known dynamic priors in context of mixture models (e.g. the OU process)
- 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

Forthcoming Research

- Apply our findings to more complex models of mixed membership, especially dynamic topic models [4]
- Leads to scalable variational inference scheme for this model class *and* to possibility of incorporating broader range of prior assumptions on topic diffusion
- Place priors on hyperparameters to capture properties of dynamic models, e.g. jumps, heteroscedasticity or stochastic volatility

References

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

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

[3] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

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

Acknowledgements

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 i:DSem research program, project PREDICT (031L0023A).