Less inefficient inference in Nonparametric Bayesian models

David Knowles

UNIVERSITY OF CAMBRIDGE

January 22, 2009

Outline

Motivation

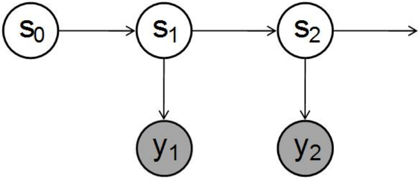
Nonparametric models have the potential to avoid overfitting or underfitting by learning appropriate model capacity

but

Many new inference algorithms struggle to outperform Gibbs sampling

Outline

Hidden Markov Model



- Hidden Markov Models have the form:

$$p(s, y|\pi_0, \pi, \phi, K) = \prod_{t=1}^{T} p(s_t|s_{t-1})p(y_t|s_t)$$

  where s is the state trajectory and y is a vector of observations through time.
- Prior on row π_k of transition matrix:

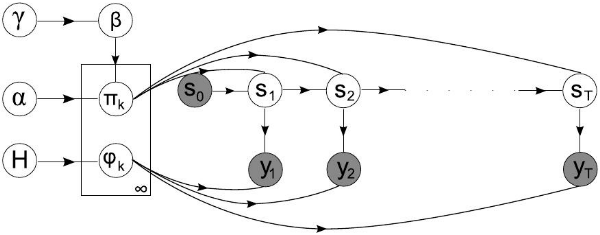
$$\pi_k \sim Dirichlet(\alpha\beta)$$

$$\beta \sim Dirichlet(\gamma/K, \ldots, \gamma/K)$$

The infinite HMM

Take the limit as $K \rightarrow \infty$

$$
\begin{aligned}
&\beta \sim GEM(\gamma) \\
&\pi_k|\beta \sim DP(\alpha, \beta) \\
&\phi_k \sim H \\
&s_t|s_{t-1} \sim Multinomial(\pi_{s_{t-1}}) \\
&y_t|s_t \sim F(\phi_{s_t})
\end{aligned}
$$

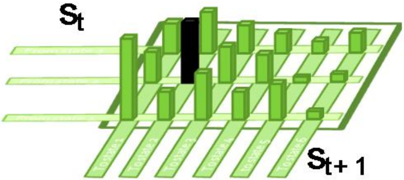


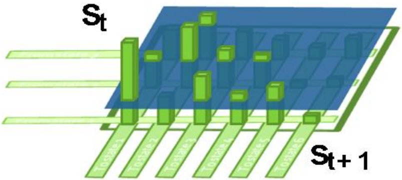
Gibbs sampling

- Integrate out π, ϕ.
- To sample state trajectories: for $t = 1..T$ compute $p(s_t | s_{-t}, \beta, y, \alpha, H)$. Some probability of transitioning into a previously unseen state.
- Very slow mixing because of strong correlations between time points

Beam sampling

Adaptive truncation with convergence to true posterior maintained
Introduce auxiliary variables $u_t \sim Uniform(0, \pi_{s_{t-1}s_t}) \forall t = 1..T$





Beam sampling

To sample state trajectories:

- Forward sweep becomes a *finite* sum:

$$p(s_t|y_{1:t}, u_{1:t}) \propto p(y_t|s_t) \sum_{s_{t-1}: u_t < \pi_{s_{t-1}s_t}} p(s_{t-1}|y_{1:t-1}, u_{1:t-1})$$

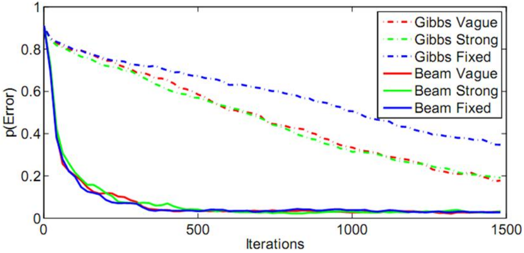
- Backwards sampling

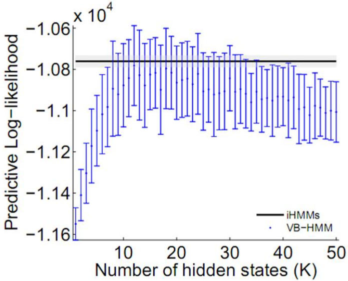
$$s_T \sim p(s_T|y_{1:T}, u_{1:T})$$

For $t = T - 1..1$

$$\begin{aligned} s_t|s_{t+1} &\sim p(s_t|s_{t+1}, y_{1:T}, u_{1:T}) \\ &\propto p(s_t|y_{1:t}, u_{1:t}) p(s_{t+1}|s_t, u_{t+1}) \end{aligned}$$

Results


Gibbs Vague
Gibbs Strong
Gibbs Fixed
Beam Vague
Beam Strong
Beam Fixed
p(Error)
Iterations
0
500
1000
1500
0
0.2
0.4
0.6
0.8
1


x 10^4
Predictive Log−likelihood
−1.06
−1.08
−1.1
−1.12
−1.14
−1.16
iHMMs
VB−HMM
10
20
30
40
50
Number of hidden states (K)

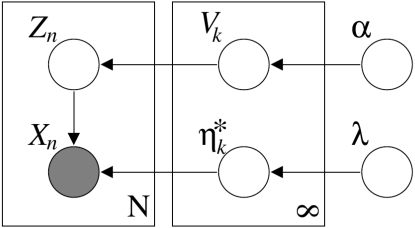
Outline

Variational Inference for DP Mixtures (Blei, Jordan 2006)

- Observations X_n, indicator variables Z_n, cluster parameters η_k
- Use the stick breaking construction for the DP:

$$v_i|\alpha \sim Beta(1, \alpha)$$

$$\pi_i|v = v_i \prod_{j=1}^{i-1}(1 - v_j)$$



Variational Inference for DP Mixtures (Blei, Jordan 2006)

- Mean field variational approximation:

$$q(v, \theta, z) = \prod_{t=1}^{T-1} q(v_t) \prod_{t=1}^{T} q(\eta_t) \prod_{n=1}^{N} q(z_n)$$

- And truncate: $q(v_T = 1) = 1$

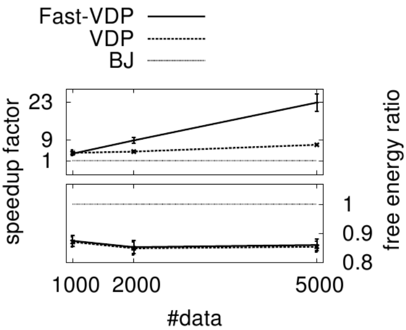
Unfortunately...

- Outperformed by Gibbs sampling (although does converge faster)
- Successive variational families are not nested, so the approximation may get *worse* increasing T to T+1

Accelerated Variational Dirichlet Process Mixtures (Kurihana, Vlassis, Welling 2006)

- Idea: instead of truncating the stick breaking construction, fix the variational distribution of all components for $k > K$ at their prior
- Still have to evaluate an infinite sum, but tractable
- Show improved performance
- (Also improve performance by cutting up sample space with kd-trees, but not really an idea that extends to other models...)

Worst plot ever?



Outline

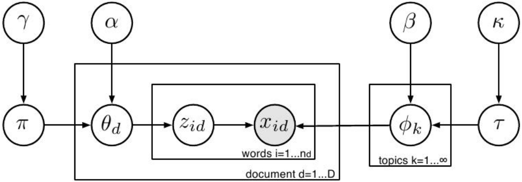
Collapsed Variational Inference for HDP (Teh, Kurihara, Welling 2008)

A nonparametric model for LDA

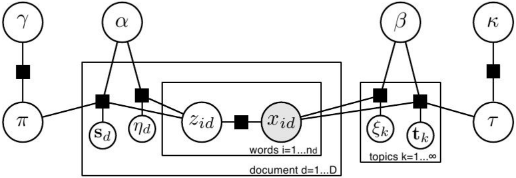
$$\begin{aligned} x_{id}|z_{id}, \phi_{z_{id}} &\sim Mult(\phi_{z_{id}}) \\ z_{id}|\theta_d &\sim Mult(\theta_d) \\ \theta_d|\pi &\sim Dir(\alpha\pi) \\ \phi_k|\tau &\sim Dir(\beta\tau) \\ v_i|\alpha &\sim Beta(1, \alpha) \\ \pi_i|v &= v_i \prod_{j=1}^{i-1}(1 - v_j) \end{aligned}$$

Graphical model

Graphical model for HDP topic model:



Factor graph including auxiliary variables

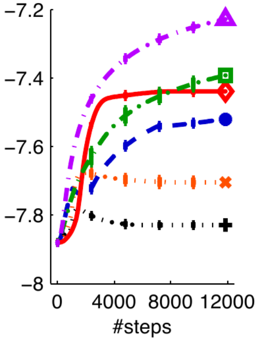


Different truncation scheme

- Idea: Assume $q(z_{id} > K) = 0$ for all i and d.
- Observations have no effect on v_k or ϕ_k for all $k > K$, so marginalise these out
- Simpler than tying to the prior but variational families at successive truncation levels are nested

Results

Log probability of test data:



- Outperforms parametric LDA
- Still outperformed by collapsed Gibbs sampling for HDP

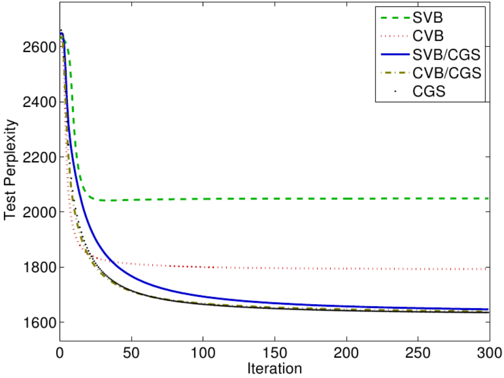
Outline

Hybrid variational/Gibbs Collapsed Inference in Topic Models (Welling, Teh, Kappen 2008)

- Idea: Combine sampling and variational approximation in a principled way
- Divide dataset of word counts per document into a set with counts $\leq r$ (call this S^{GB}) and $> r$ (call this S^{VB})
- Gibbs sampling for the S^{GB}
- Variational approximation for S^{VB}
- Assume factorised across division and combine in a principled way
- Stochastically maximises the variational bound

Hybrid variational/Gibbs Collapsed Inference in Topic Models (Welling, Teh, Kappen 2008)

A lot of work... and now we can rival collapsed Gibbs sampling! With $r = 1$



Outline

Conclusions

- Significantly outperforming Gibbs sampling is hard!
- "Slicing up" nonparametric models ala beam sampling can be very effective
- There is significant interest in getting variational approximations to work in nonparametric models
- Truncation strategies, collapsing and auxiliary variables are important
- Hybrid sampling/variational methods may be useful but generalisation to continuous variables not yet clear