Variational Inference and Big Data

John Paisley

Department of Electrical Engineering
Columbia University

Joint work with Matt Hoffman, Chong Wang, David Blei and Michael Jordan

Model structure

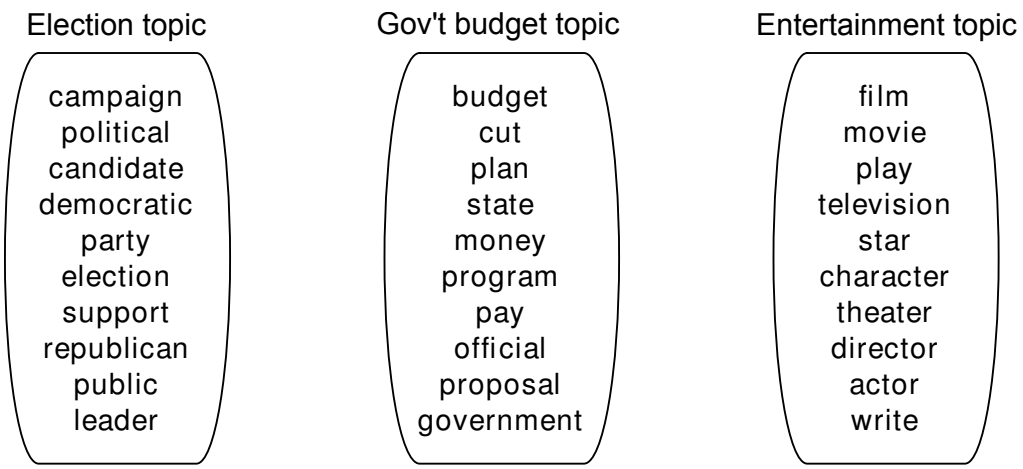
The class of models we are interested in involves observations, global hidden variables, local hidden variables, and fixed parameters.

1. Global variables: β
2. Local variables: $z_1, \ldots, z_n$
3. Other fixed parameters: α

The local variables govern distributions on the data $x_1, \ldots, x_n$.

Given global variables, what happens locally is independent of each other.

Example: Topic models



- A core idea is of a "topic"— a probability distribution on words.
- Words in a document generated from some combination of topics.
- Text models differ mainly in how they put together topics to build up larger models appropriate for the specific problem goals.

Example topic model: Latent Dirichlet allocation (LDA)

- Draw topics (distributions on words): $\beta_k \sim \text{Dirichlet}, k = 1, \ldots, K$.
- Draw distribution on topics for each document: $\theta_i \sim \text{Dirichlet}$.
- For word n in document i,
  - Sample topic indicator $c_{ni} \sim \text{Discrete}(\theta_i)$,
  - Sample word value $x_{ni} \sim \text{Discrete}(\beta_{c_{ni}})$.

The global variables are $\beta = \{\beta_1, \ldots, \beta_K\}$.

The local variables are $z_i = \{\theta_i, c_{1i}, \ldots, c_{ni}\}$.

Example topics

The New York Times

music band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach	show film television movie series says life man character know
theater play production show stage street broadway director musical directed	clinton bush campaign gore political republican dole presidential senator house	stock market percent fund investors funds companies stocks investment trading	restaurant sauce menu food dishes street dining dinner chicken served	budget tax governor county mayor billion taxes plan legislature fiscal

Model structure

A property these models share is that the joint distribution factorizes over the local variables,

$$p(x, \beta, z|\alpha) = p(\beta|\alpha) \prod_{i=1}^{n} p(x_i, z_i|\beta, \alpha).$$

Given the global variables, the local variables and corresponding data are conditionally independent.

Our goal is to approximate the posterior distribution of the hidden variables given the observations,

$$p(\beta, z|x, \alpha) = \frac{p(x, \beta, z|\alpha)}{\int_{\beta,z} p(x, \beta, z|\alpha) d\beta dz}.$$

Variational inference

Mean-field variational Bayes is a deterministic inference method for approximating the posterior distribution of a Bayesian model.

1. Define a factorized distribution to approximate the posterior,

$$p(\beta, z|x) \approx q(\beta) \prod_i q_i(z_i).$$

2. Lower bound the marginal likelihood,

$$\begin{aligned} \ln p(x|\alpha) &= \ln \int_{\beta,z} p(x, \beta, z|\alpha) d\beta dz \\ &\geq \int_{\beta,z} q(\beta) \prod_i q_i(z_i) \ln \left\{ \frac{p(x, \beta, z|\alpha)}{q(\beta) \prod_i q_i(z_i)} \right\} d\beta dz \\ &:= \mathcal{L}. \end{aligned}$$

3. Maximize $\mathcal{L}$ with respect to parameters of each q. This is equivalent to minimizing the KL divergence between the full posterior and q.

Conjugate models

Conditional independence in the likelihood:

$$p(w_1, \ldots, w_n|\eta) = \prod_{i=1}^{n} p(w_i|\eta) = \left[\prod_{j=1}^{n} h(w_j)\right] \exp\left\{\eta^T \sum_j t(w_j) - nA(\eta)\right\}$$

A conjugate prior for η:

$$p(\eta|\chi, \nu) = f(\chi, \nu) \exp\{\eta^T \chi - \nu A(\eta)\}$$

Through Bayes rule the posterior is in the same family:

$$p(\eta|\chi', \nu') = f(\chi', \nu') \exp\{\eta^T \chi' - \nu' A(\eta)\},$$

with $\chi' = \chi + \sum_j t(w_j)$ and $\nu' = \nu + n$.

The variational take on conjugacy

Define the q distribution of η as

$$q(\eta|\chi', \nu') = f(\chi', \nu') \exp\{\eta^T \chi' - \nu' A(\eta)\}.$$

Optimize the variational objective wrt parameters χ' and ν'.

By differentiating variational objective with respect to $[\chi', \nu']^T$, we find

$$\nabla \mathcal{L} = - \begin{bmatrix} \frac{\partial^2 \ln f(\chi', \nu')}{\partial \chi' \partial \chi'^T} & \frac{\partial^2 \ln f(\chi', \nu')}{\partial \chi' \partial \nu'} \\ \frac{\partial^2 \ln f(\chi', \nu')}{\partial \nu' \partial \chi'^T} & \frac{\partial^2 \ln f(\chi', \nu')}{\partial \nu'^2} \end{bmatrix} \begin{bmatrix} \chi + \sum_j \mathbb{E}_q t(w_j) - \chi' \\ \nu + n - \nu' \end{bmatrix}.$$

Setting this to zero we can read off the update to $q(\eta|\chi', \nu')$.

General form of batch variational inference

For each parameter η_k in the conjugate exponential model:

- Define a distribution $q(\eta_k)$ in same family as the prior.
- For each η_k, update the parameters of $q(\eta_k)$ with the expected sufficient statistics wrt current $q(\eta_j)$, $j \neq k$.

Gibbs sampling: Sample from this distribution using sufficient statistics of the current sample.

Writing out the variational objective

The variational objective for our model framework is

$$\mathcal{L} = \sum_{i=1}^{n} \mathbb{E}_q \ln\{p(x_i, z_i|\beta, \alpha)/q(z_i)\} + \mathbb{E}_q \ln p(\beta|\alpha) - \mathbb{E}_q \ln q(\beta)$$

Batch algorithm:

1. For each i, optimize $q(z_i)$
2. Optimize $q(\beta)$
3. Repeat

When n is huge and a non-trivial amount of work is needed to optimize each $q(z_i)$, Step 1 can take a very long time.

Stochastic variational inference

Because the likelihood factorizes, the objective function splits into a sum over local variable terms,

$$\mathcal{L} = \underbrace{\sum_{i=1}^{n} \mathbb{E}_q \ln\{p(x_i, z_i|\beta, \alpha)/q(z_i)\}}_{\text{local variables for } q(z_i)} + \mathbb{E}_q \ln p(\beta|\alpha) - \mathbb{E}_q \ln q(\beta).$$

This suggests we can use stochastic optimization to maximize $\mathcal{L}$.

Stochastic variational inference

Create a new objective function by sampling and scaling,

$$\mathcal{L}_t = \underbrace{\sum_{i \in C_t} \frac{n}{|C_t|} \mathbb{E}_q \ln\{p(x_i, z_i|\beta, \alpha)/q(z_i)\}}_{\text{local variables for subset of } q(z_i)} + \mathbb{E}_q \ln p(\beta|\alpha) - \mathbb{E}_q \ln q(\beta).$$

We can show that $\mathbb{E}[\mathcal{L}_t] = \mathcal{L}$.

Stochastic algorithm:

1. Select a small subset of $q(z_i)$ at random and optimize
2. Scale up the impact of the subset by $n/|C_t|$
3. Take a step in direction of *natural gradient* for $q(\beta)$
4. Repeat

Stochastic variational inference

Take gradient step for the global $q(\beta)$:

$$[\chi', \nu']^T = [\chi'_{\text{old}}, \nu'_{\text{old}}]^T + \rho_t B \nabla \mathcal{L}_t$$

In this case, the gradient is:

$$\nabla \mathcal{L}_t = - \begin{bmatrix} \frac{\partial^2 \ln f(\chi', \nu')}{\partial \chi' \partial \chi'^T} & \frac{\partial^2 \ln f(\chi', \nu')}{\partial \chi' \partial \nu'} \\ \frac{\partial^2 \ln f(\chi', \nu')}{\partial \nu' \partial \chi'^T} & \frac{\partial^2 \ln f(\chi', \nu')}{\partial \nu'^2} \end{bmatrix} \begin{bmatrix} \chi + \frac{n}{|C_t|} \sum_{j \in C_t} \mathbb{E}_q t(w_j) - \chi' \\ \nu + n - \nu' \end{bmatrix}.$$

Simplifying inference

Set B to inverse Fisher information of $q(\eta)$ and the left matrix cancels.

$$\chi' = (1-\rho_t)\chi'_{\text{old}} + \rho_t \left(\chi + \frac{n}{|C_t|} \sum_{j \in C_t} \mathbb{E}_q t(w_j) \right),$$
$$\nu' = (1-\rho_t)\nu'_{\text{old}} + \rho_t \left(\nu + n \right).$$

If the step size ρ_t is such that $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$, this approach converges to a local optimal of $\mathcal{L}$.

Stochastic variational inference

A Dirichlet-multinomial example:

$$\theta \sim \text{Dirichlet}(\chi), \quad z_i \overset{iid}{\sim} \text{Multinomial}(\theta), \quad q(\theta) = \text{Dirichlet}(\chi')$$

Where z_i is a latent indicator variable in a mixture model.

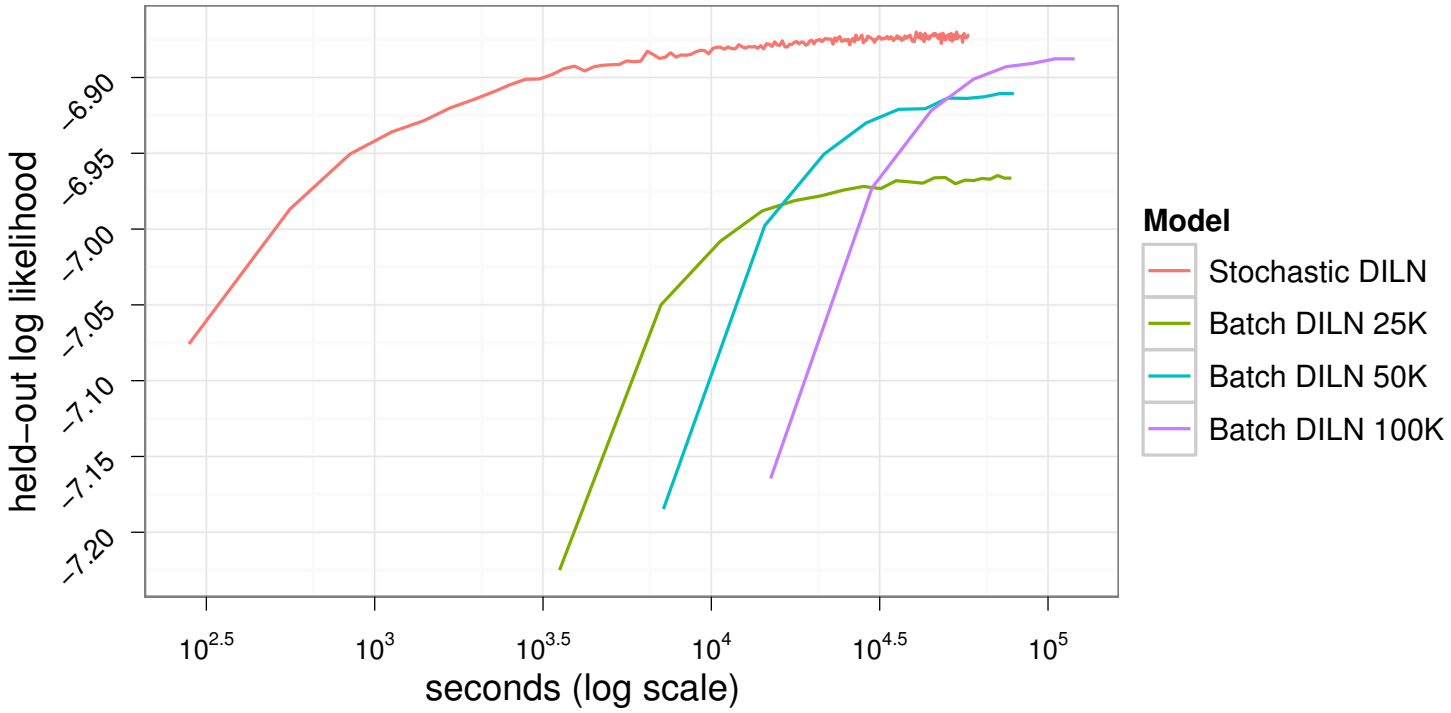
Batch update: $\chi' = \chi + \sum_i \mathbb{E}_q z_i$

- Requires update of each $q(z_i)$.

Stochastic update: $\chi'_t = (1 - \rho_t)\chi'_{\text{old}} + \rho_t(\chi + \frac{n}{|C_t|} \sum_{i \in C_t} \mathbb{E}_q z_i)$

- Requires update of $q(z_i)$ only for $i \in C_t$.

Batch vs stochastic inference



- Stochastic vs batch inference using 350K articles from *Nature*.

References

Main:

1. M. Hoffman, D. Blei, C. Wang and J. Paisley (2013). Stochastic variational inference. *Journal of Machine Learning Research,* vol. 14, pp. 1303-1347.

Two applications:

1. J. Paisley, C. Wang, D. Blei and M. Jordan (2012). Nested hierarchical Dirichlet processes. arXiv:1210.6738.

2. J. Paisley, C. Wang and D. Blei (2012). The discrete infinite logistic normal distribution. *Bayesian Analysis*, vol. 7, no. 2, pp. 235-272.