The Population Posterior and Bayesian Modeling on Streams

(McInerney, Ranganath, Blei 2015)

This paper proposes a framework for doing Bayesian modelling on streams. It could be applied to anything in theory (in the paper they did this for LDA and Dirichlet process mixtures) ,but in this note we focus on drawing analogies to the standard LDA.

The problem with Bayesian models like LDA is that it assume several things about the data (model correctness, exchangeability..etc) I think the principle motivation for this paper is the idea that streaming violates the exchangeability assumption. The original LDA assumes that order does not matter for words in a topic or how the documents in a corpus are ordered. This is the **bag of words** assumption which is completely reasonable when our data is static. But for streaming, we obviously want some way to distinguish data from yesterday versus data from 10 years ago. So this exchageablilty assumption cannot hold anymore.

Population Posterior

We have β which is a **global** latent variable that governs any point, and z_i a local hidden variable that only governs the ith datapoint.

This model is given by the joint:

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

This equation kind of make sense because since β is global, it stands alone and doesn't need to be incorporated in the product over all datapoints. And since β is global, both the datapoint x_i and the local hidden variable z_i kind of depend on it. It also follows directly if you just apply the chain rule to decompose the joint into conditional probabilities.

To account for the fact that our data is changing, we define a new quantity $\mathbf{X}$. $\mathbf{X}$ is the population that's drawn from the distribution F_α of size α. We consider a latent variable model of α data points that we are streaming in.

Our **population posterior** is defined as the the expectation of our posterior. In standard LDA, we were interested in the posterior $p(\theta, z|w)$, i.e. what is the topic distribution and topic assignment given that we saw a

word. Simmilarly, here we are interested in the posterior $p_i(\beta, z|\mathbf{X})$. Note the difference here, the reason why we take the expectation of the posterior w.r.t the distribution is that we don't have one posterior, we have many population posteriors! As the data streams in, the F_α distribution that we draw $\mathbf{X}$ from changes, so X will change too. We end up having many population distributions that we need to average or aggregate over.

$$\mathbf{E}_{F_\alpha}[p(\beta, z|X)] = \mathbf{E}_{F_\alpha}\left[\frac{p(\beta, z, X)}{p(X)}\right]$$

Variational Inference for Population Posterior (F-ELBO)

In the variational Bayes, we minimize the KL divergence between our lowerbound function q and posterior to maximize the ELBO. By analogy in **population variational Bayes**, our goal is to minimize the KL divergence between our lowerbound function $q(\beta, z)$ and $E_{F_\alpha}[p(\beta, z|X)]$ to maximize the F-ELBO. So through the expectation, our objective function (F-ELBO) is a function of the population distribution F_α rather than a fixed dataset x(and also a function of expectation of q).

Again here, we have our guess function q (which is defined very similarly as in LDA with notational differences):

$$q(\beta, \mathbf{z}) = q(\beta|\lambda)\prod_{i=1}^{\alpha} q(z_i|phi_i)$$

where λ is the global parameters and ϕ_i are the local parameters.

To derive the F-ELBO (see Appendix A), we start off with the original relationship between the ELBO with the KL divergence.

$$logp(x) = D(q(\beta, z)||p(z|x)) + \overbrace{E_q[logp(\beta, z, x) - logq(\beta, z)]}^{ELBO}$$

Now, rather than a fixed dataset x we have X, so we also need to now take the expectation w.r.t F_α as done in Eq.2. So taking $\mathrm{E}_{F_\alpha}[\ldots]$ on both sides, then doing the regular Jensen's inequality stuff to rearrange to a KL divergence form, we get F-ELBO as:

$$\mathscr{L}(\lambda, \phi; F_\alpha) = \mathrm{E}_{F_\alpha}\left[\mathrm{E}_q\left[logp(\beta) - logq(\beta|\lambda) + \sum_{i=1}^{\alpha} logp(X_i, Z_i|\beta) - logq(Z_i)\right]\right]$$

Experimental Results:

In order to evaluate various inference methods, we look at the held-out likelihood. This is like doing cross-validation for topic models. What you do is you hold out a set of documents $\mathbf{w_d}$ of unseen document as test set, to compute the log-likelihood:

$$\mathscr{L}(w) = logp(w|\Phi, \alpha) = \sum_{d} logp(w_d|\Phi, \alpha)$$

. It turns out that evaluating this quantity is intractable, but there are many common estimators that could be used to approximate this (Wallach09). The key thing to note here is just the higher the held-out likelihood, the better the model.

They ran 3 experiments comparing:

- Inference methods (SVI, VB, SVB)
- Application to LDA and Dirchlet process mixture (we ignore this comparison, only care about LDA results)
- varying α (number of datapoints in streams)

References:

- David Blei's course notes is pretty good for understanding the basics of variational bayes, exponential families and how they are used in conditional conjugate models
- Good list of papers related to topic modelling and their 1-sentence summaries.