Summary 3: Correlated Mixture Models

Susan Cherry

This paper introduces Correlated Topic Models (CTM), which can be used to model the mixture of topics in a set of documents. Technically, a topic model is "generative probabilistic model that uses a small number of distributions over a vocabulary to describe a document collection." The simplest topic model is the Latent Dirichlet Allocation (LDA), which CTM is built off. One of the draw backs of the LDA is that it doesn't allow for correlation between topics. For example, an scholar searching a body of documents for one topic would probably be interested in knowing if that topic is correlated with other topics (genetics is probably correlated with health). The inability of the LDA to model correlation comes from the independence assumption of the Dirichlet Distribution.

The CMT allows for correlation by introducing the logistic normal distribution instead of the Dirichlet, which is more flexible. The covariance matrix of the logistic normal distribution models the correlation. Unfortunately, the logistic normal is not conjugate to the multinomial distribution, so posterior inference is more difficult than it is using the LDA. The authors use Variational Expectation Maximization.

The paper concludes by discussing empirical results comparing the LDA and CMT, using over 16,000 articles from the magazine *Science*. In all cases, the CMT performed better. The CMT provided a better fit and fitted more topics the the LDA. Also, the CMT did a better job of predicting remaining words when shown part of a document. It had less uncertainty about the remaining words than LDA (reduced perplexity by 200 words).

The main takeaways from the paper were: 1) CMT allows us to model correlation between topics, which is not possible with LDA. 2) The drawback to CMT is that the logistic normal is not conjugate, so posterior inference is more difficult.