Summary of “Correlated Topic Models”

Matt Dickenson `mcd31`

STA571/CS590.01 Due: 22 January, 2014

Correlated topic modeling (CTM) has several advantages over earlier methods of topical analysis, such as LDA. The main difference between these two models is that CTM does not assume independence between topics (as LDA does, through its reliance on the Dirichlet distribution). This means that CTM can borrow information across topics, and needs to observe less of a given document than an LDA model would to achieve the same predictive accuracy. CTM is also helpful for visual exporation of unstructured datasets.

These benefits are not without cost, however. Most importantly, the logistic normal distribution used in CTM is not conjugate with the multinomial distribution (used for topics), so analytical methods are intractable. The goal in inference with CTM models is to minimize the Kullback-Leibler divergence between the approximate and true posterior. MCMC is impractical for this, and so expectation-maximization (EM) is used. Unfortunately EM does not have the same asymptotic guarantees as MCMC, so the steps are iterated until the change in the likelihood between iterations is arbitrarily small.

Improvements on CTM will likely have to address shortcomings that it shares with LDA. For example, in both models the set of topics is assumed to be fixed across all the data. More advanced models for textual analytics could include topic discovery by incorporating components of change-point analysis. The potential number of topics k in each period t could be included in the model as a latent variable to be estimated. Overall, CTM is a useful modeling strategy whose predictive accuracy more than makes up for its complexity.