

Probabilistic Topic Models (Blei, 2012)

Satvik Saha (ss7228)

September 30, 2025

This article surveys probabilistic topic models, focusing on *Latent Dirichlet Allocation* (LDA), which can be described by the generative model

$$\begin{aligned}\beta_k &\sim \text{Dirichlet}(\eta), \\ \theta_i &\sim \text{Dirichlet}(\alpha), \\ z_{ij} \mid \theta_i &\sim \text{Categorical}(\theta_i), \\ w_{ij} \mid z_{ij}, \beta &\sim \text{Categorical}(\beta_{z_{ij}}),\end{aligned}$$

where $\eta \in \mathbb{R}^V$, $\alpha \in \mathbb{R}^K$. We only observe a collection of words $\{w_{ij}\}$, with w_{ij} being the i -th word (from a vocabulary of V words) from the j -th document. The goal is to discover a hidden structure of K topics $\{\beta_k\}$, with each topic β_k being a distribution over words. Any given document (which itself can be viewed as a distribution over words) is assumed to be a mixture of these topics, with proportions θ_i ; this makes LDA a kind of *mixed membership model* as opposed to a simple mixture model. The hyperparameter α controls the sparsity of θ_i ; sparser θ_i meaning that fewer topics dominate in the corresponding document.

LDA can be viewed as a kind of factorization model: the observed document-word distributions are factored into document-topic distributions θ and topic-word distributions β . A non-probabilistic version would involve matrix factorization, such as in *Nonnegative Matrix Factorization* (NMF).

LDA can also be viewed as a kind of dimensionality reduction technique: the inferred document-topic distribution $\theta_i \in \Delta^{K-1}$ is a summary of the underlying document-word generating distribution (which lives in Δ^{V-1}) when $K \ll V$.

A crude way of incorporating metadata (publishing information, location, etc) $\{x_i\}$ for each document might be to draw $\theta_i \mid \alpha_i \sim \text{Dirichlet}(\alpha_i)$, with α_i itself regressed on x_i in some way. This might help explore how different factors in x_i push documents towards particular topics.

References

David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.