

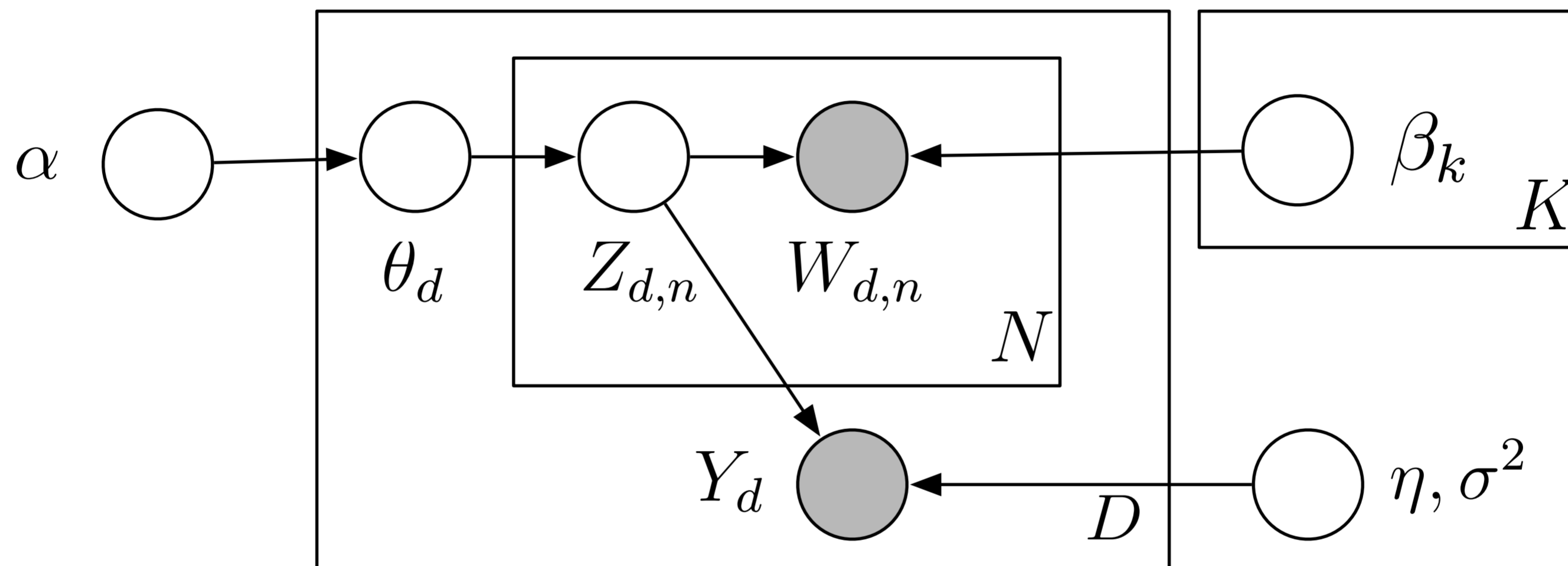
Supervised Topic Models

David M. Blei

Department of Computer Science
Princeton University

Jon D. McAuliffe

Department of Statistics
University of Pennsylvania



We introduce supervised latent Dirichlet allocation (sLDA), a statistical model of labelled documents. The model accommodates a variety of response types. We derive a maximum-likelihood procedure for parameter estimation, which relies on variational approximations to handle intractable posterior expectations. Prediction problems motivate this research: we use the fitted model to predict response values for new documents. We test sLDA on two real-world problems: movie ratings predicted from reviews, and web page popularity predicted from text descriptions. We illustrate the benefits of sLDA versus modern regularized regression, as well as versus an unsupervised LDA analysis followed by a separate regression.