STM

The Structural Topic Model (STM) enables the discovery of topics and the estimation of their relationship to document metadata by providing a general way to corporate corpus structure or document metadata into the standard topics model.

The STM model accommodates corpus structure through document-level covariates affecting topical prevalence and/or topical content by specifying the priors as generalized linear models.

The STM model specifies two design matrices of covariates for topic prevalence and topical content where each row defines a vector of covariates for a given document. X represents the topic prevalence matrix while Z represents the topical content matrix.

The topic prevalence component allows the expected document-topic proportions to vary by covariates X rather than arising from a single shared prior. We model the mean vector of the Logistic Normal as a simple linear model model such that μi = Xiγi giving γ a regularizing prior to avoid over fitting. Intuitively this takes the form of a normal multivariate linear model with shared covariance parameters. For topic content, it used Z instead. The κ is a given sparsity inducing priors so that topic and covariate effects represent sparse deviations from the corpus-wide empirical word frequency.