Theoretical Considerations

The Structural Topic Model (STM) is a topic model which extends classical topic models such as Latent Dirichlet Allocation (LDA) by incorporating information of covariates. Topic models can be used to infer topics from a large text corpus grouped into documents. In topic modelling it is assumed that this corpus is generated from a small number of distributions over words, the topics. The proportions of these topics are document-specific. In contrast to simpler topic models such as LDA, the STM relates topic proportions to document-level covariates. Furthermore, each distribution over words, i.e. each topic, can vary for different documents dependent on the covariate values of this document.

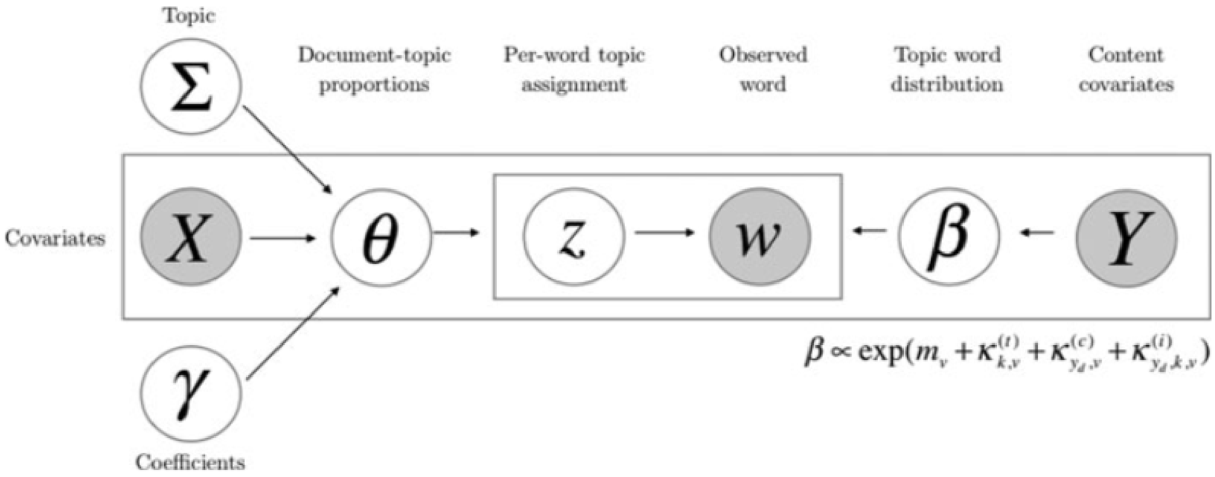
Topic
Σ
Document-topic proportions
Per-word topic assignment
Observed word
Topic word distribution
Content covariates
Covariates
X
θ
z
w
β
Y
$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$
γ
Coefficients

Figure 1: Graphical model representation of the STM

Posterior

$$\begin{aligned} p(\eta, z, \kappa, \gamma, \Sigma | w, X, Y) &\propto \underbrace{p(w|\eta, z, \kappa, \gamma, \Sigma, X, Y)}_{=p(w|z,\kappa,Y)} p(\eta, z, \kappa, \gamma, \Sigma | X, Y) \\ &\propto p(w|z, \kappa, Y) p(z|\eta) p(\eta|\gamma, \Sigma, X) p(\kappa) p(\gamma, \Sigma) \\ &\propto \Big\{ \prod_{d=1}^{D} p(\eta_d|\gamma, \Sigma, X) \Big(\prod_{n=1}^{N} p(w_n|\beta_{d,k=z_{d,n}}) p(z_{d,n}|\theta_d) \Big) \Big\} \prod p(\kappa) \prod p(\gamma), \end{aligned}$$

where $\beta_{d,k,\nu} \propto \exp(m_\nu + \kappa^{(t)}_{k,\nu} + \kappa^{(c)}_{y_d,\nu} + \kappa^{(i)}_{y_d,k,\nu})$ and $\theta_d := \text{softmax}(\eta_d)$.