analysis_patrick

Method of composition

Let $\theta_{(k)} \in \mathbb{R}^D$ denote the proportion of the k-th topic for all D documents. Suppose that we want to perform a regression of these topic proportions $\theta_{(k)}$ on a subset $\tilde{X} \in \mathbb{R}^{D \times \tilde{P}}$ of prevalence covariates X. The true topic proportions are unknown, but the STM produces an estimate of the approximate posterior of $p(\theta_{(k)}|\Gamma, \Sigma, X)$. A naïve approach would be to regress the estimated mode or mean of the approximate posterior on $\tilde{X}$. However, this approach neglects much of the information contained in the approximate posterior distribution of $\theta_{(k)}$. Instead, sampling $\theta_{(k)}$ from this distribution, performing a regression for each sampled $\theta_{(k)}$ on $\tilde{X}$, and then sampling from the estimated distributions of regression coefficients, provides an i.i.d. sample from the marginal posterior distribution of regression coefficients.

Formally, let $p(\xi|\theta_{(k)}, \tilde{X})$ denote the distribution of regression coefficients ξ resulting from estimating the linear model $\theta_{(k)} = \tilde{X}\xi + \epsilon$, where $\epsilon \in \mathbb{R}^D$ are uncorrelated normally distributed errors with mean zero. Let further $\tilde{p}(\theta_{(k)}|\Gamma, \Sigma, X)$ denote the approximate posterior of $p(\theta_{(k)}|\Gamma, \Sigma, X)$.

Repeat m times:

1. Draw $\theta^*_{(k)} \sim \tilde{p}(\theta_{(k)}|\Gamma, \Sigma, X)$.
2. Draw $\xi^* \sim p(\xi|\theta^*_{(k)}, \tilde{X})$.

Then, $\xi^*_1, \ldots, \xi^*_m$ is an i.i.d. sample from the marginal posterior

$$\tilde{p}(\xi|\Gamma, \Sigma, X) := \int_{\theta_{(k)}} p(\xi|\theta_{(k)}, \tilde{X})\tilde{p}(\theta_{(k)}|\Gamma, \Sigma, X)\mathrm{d}\theta_{(k)} = \int_{\theta_{(k)}} \tilde{p}(\xi, \theta_{(k)}|\Gamma, \Sigma, X)\mathrm{d}\theta_{(k)},$$

where $\tilde{p}(\xi, \theta_{(k)}|\Gamma, \Sigma, X)$ denotes the joint distribution of $\xi|\theta_{(k)}, \tilde{X}$ and the approximate posterior of $\theta_{(k)}$. Thus, it has been integrated over the latent variable $\theta_{(k)}$, incorporating uncertainty about $\theta_{(k)}$, when determining ξ.