Method of composition

Let $\theta_{(k)} \in [0,1]^D$ denote the proportions 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 $q(\theta_{(k)}|\Gamma, \Sigma, X)$ of $\theta_{(k)}$, where $\Gamma := \Gamma(w, X, Y)$ and $\Sigma := \Sigma(w, X, Y)$. A naïve approach would be to regress the estimated mode of the approximate posterior distribution on $\tilde{X}$. However, this approach neglects much of the information contained in the distribution. Instead, sampling $\theta^*_{(k)}$ from the posterior 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. This procedure is known as the method of composition in the social sciences (Tanner 2012, 52)

Formally, let ξ denote the regression coefficients from a regression of $\theta_{(k)}$ on $\tilde{X}$, and let $q(\xi|\theta_{(k)}, \tilde{X})$ be the approximate posterior distribution of these coefficients, i.e. given design matrix $\tilde{X}$ and response $\theta_{(k)}$. To adequate model the topic proportions we perform a beta regression, because the sampled proportions are restricted to the interval $(0, 1)$. More information on why the beta regression is useful in such a scenario can be found in Ferrari and Cribari-Neto (2004). In case of a beta regression $q(\xi|\theta_{(k)}, \tilde{X})$ is a normal distribution (see e.g. Ferrari and Cribari-Neto (2004), p. 17). In contrast to our approach, the R package *stm* implements a simple OLS regression. As expected, using this framework we frequently observed predicted proportions outside of $(0, 1)$. Moreover, credible intervals are not informative, due to violated model assumptions.

The method of composition can now be described by repeating the following process m times:

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

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

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

where $q(\xi, \theta_{(k)}|\Gamma, \Sigma, X) := q(\xi|\theta_{(k)}, \tilde{X}) q(\theta_{(k)}|\Gamma, \Sigma, X)$. Thus, it has been integrated over $\theta_{(k)}$, which allows to incorporate uncertainty about $\theta_{(k)}$, when determining ξ.

Ferrari, Silvia, and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7). Taylor & Francis: 799–815.

Tanner, Martin A. 2012. *Tools for Statistical Inference*. Springer.