# 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 or mean 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.

Formally, let $\xi$ denote the regression coefficients and $q(\xi|\theta_{(k)},\tilde{X})$ the distribution of these coefficients, given design matrix $\tilde{X}$ and response $\theta_{(k)}$.

Repeat $m$ times:

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

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

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 $\xi$.