Description of topic models

LDA

The generative process of latent Dirichlet allocation (LDA) is defined as:

1. draw $\phi_k \sim Dir(\alpha)$, for each signature $k$, where $Dir(\alpha)$ is a symmetric Dirichlet distribution with parameter $\alpha$
2. draw $\theta_d \sim Dir(\beta)$, for each sample $d$
3. draw signature assignment, $z_{dn} \sim Cat(\theta_d)$, for each mutation $n$, in each sample $d$, where $Cat(\theta)$ is a categorical distribution with probabilities $\theta$
4. draw $x_{dn} \sim Cat(\phi_{z_{dn}})$, for each mutation $n$, in each sample $d$

Given parameters $\alpha$ and $\beta$, the joint distribution of mutations $X$, mutation-signature assignments $Z$, signature mixtures $\theta$, and signatures $\phi$ is given by:

\[
p(X, Z, \theta, \phi | \alpha, \beta) = \prod_{k=1}^{K} p(\phi_k | \alpha) \times \prod_{d=1}^{D} p(\theta_d | \beta) \prod_{n=1}^{N_d} p(z_{dn} | \theta_d) p(x_{dn} | z_{dn}, \phi)
\]

where $K$ is the number of signatures, $D$ is the number of samples, $N_d$ is the number of mutations in a sample, $d$.

CTM

The correlated topic model (CTM) modifies LDA by replacing the $\theta$ variable with a $K$-dimensional Normally-distributed variable $\eta$. The Gaussian covariance matrix captures signature correlations across samples. The generative process is:

1. draw $\phi_k \sim Dir(\alpha)$, for each signature $k$
2. draw $\eta_d \sim N(\mu, \Sigma)$, for each sample $d$, where $N(\mu, \Sigma)$ is the multivariate Gaussian distribution with mean $\mu$ and covariance $\Sigma$
3. draw signature assignment, $z_{dn} \sim Cat(f(\eta_{d1..K}))$, for each mutation $n$, in each sample $d$, where $f(\eta_{dk}) = \frac{\exp(\eta_{dk})}{\sum_{k'}^{K} \exp(\eta_{dk'})}$ transforms $\eta_{d1..K}$ to a valid probability distribution
4. draw $x_{dn} \sim Cat(\phi_{z_{dn}})$, for each mutation $n$, in each sample $d$

The joint probability then becomes:

\[
p(X, Z, \eta, \phi | \alpha, \mu, \Sigma) = \prod_{k=1}^{K} p(\phi_k | \alpha) \times \prod_{d=1}^{D} p(\eta_d | \mu, \Sigma) \prod_{n=1}^{N_d} p(z_{dn} | \eta_d) p(x_{dn} | z_{dn}, \phi)
\]
where $K$ is the number of signatures, $D$ is the number of samples, $N_d$ is the number of mutations in a sample $d$.

**MMCTM**

The multi-modal correlated topic model (MMCTM) extends the CTM further, by allowing signature inference for multiple data/mutation types (i.e. modalities) simultaneously. The generative process is as follows:

1. draw $\phi^m_k \sim \text{Dir}(\alpha^m)$, for each signature $k$, in each modality $m$

2. draw $\eta_d \sim N(\mu, \Sigma)$, for each sample $d$, where $N(\mu, \Sigma)$ is the multivariate Gaussian distribution with mean $\mu$ and covariance $\Sigma$

3. draw signature assignment, $z^m_{dn} \sim \text{Cat}(f(\eta_d^m))$, for each mutation $n$, in each sample $d$, in each modality $m$, where $\eta_d^m$ is a modality-specific subset of $\eta_d$, i.e. $\eta_d = \eta_1^d, \ldots, \eta_M^d$

4. draw $x^m_{dn} \sim \text{Cat}(\phi^m_{z^m_{dn}})$, for each mutation $n$, in each sample $d$, in each modality $m$

The joint probability is written as:

$$p(X, Z, \eta, \phi \mid \alpha, \mu, \Sigma) = \prod_{m=1}^{M} \prod_{k=1}^{K_m} p(\phi^m_k \mid \alpha^m) \times \prod_{d=1}^{D} p(\eta_d \mid \mu, \Sigma) \prod_{m=1}^{M} \prod_{n=1}^{N_m^d} p(z^m_{dn} \mid \eta_d^m)p(x^m_{dn} \mid z^m_{dn}, \phi^m)$$

where $M$ is the number of modalities, $K_m$ is the number of signatures in a modality $m$, $D$ is the number of samples, $N_m^d$ is the number of mutations in a sample modality. When using only a single modality, the MMCTM reduces to the CTM. Therefore, CTM parameters were inferred using the MMCTM implementation in this study, but with counts from a single mutation type.

**IMMCTM**

The independent-feature multi-modal correlated topic model (IMMCTM) is based on a previously described independent mutation feature model [1], as well as the MMCTM [2][3]. This model represents signatures as a collection of independent mutation features. For example, like typical SNV signatures [4] both the substitution and flanking nucleotides can be included in a signature definition; however, this model can treat the substitution and flanking nucleotides independently rather than concatenating them into a combined mutation type. This means that for signatures taking into account trinucleotide content around substitutions, there can be three features (one substitution and two flanking nucleotides) with $6 + 4 + 4 = 14$ possible feature values as opposed to the typical $6 \times 4 \times 4 = 96$ possible feature values commonly used for mutation signatures. The generative process for the IMMCTM is as follows:
1. draw $\phi_{m_1}^{m_1} \sim \text{Dir}(\alpha_{m_1}^{m_1})$, for each feature $i$, in each signature $k$, in each modality $m$

2. draw $\eta_d \sim \mathcal{N}(\mu, \Sigma)$, for each sample $d$

3. draw mutation signature assignment $z_{dn}^{m_1} \sim \text{Cat}(f(\eta_{dm}^{m_1}))$, for each mutation $n$, in each sample $d$, in each modality $m$

4. draw $x_{dni}^{m_1} \sim \text{Cat}(\phi_{z_{dni}^{m_1}})$, for each mutation feature $i$, in each mutation $n$, in each modality $m$, in each sample $d$

Then the model joint probability is

$$p(X,Z,\eta,\phi | \alpha,\mu,\Sigma) = M \prod_{m=1}^{M} K_m \prod_{k=1}^{K_m} \prod_{i=1}^{I_m} p(\phi_{m_1}^{m_1} | \alpha_{m_1}^{m_1}) \times D \prod_{d=1}^{D} p(\eta_d | \mu, \Sigma) M \prod_{m=1}^{M} N_{d}^{m_1} \prod_{n=1}^{N_{d}^{m_1}} p(z_{dn}^{m_1} | \eta_d^{m_1}) \prod_{i=1}^{I_m} p(x_{ dni}^{m_1} | \phi_{z_{dni}^{m_1}})$$

(4)

where $M$ is the number of modalities, $K_m$ is the number of signatures in a modality $m$, $D$ is the number of samples, $N_{d}^{m_1}$ is the number of mutations in a sample modality, and $I_m$ is the number of mutation features in a modality.

When using only a single modality, the IMMCTM reduces to the ICTM. Therefore, ICTM parameters were inferred using the IMMCTM implementation in this study, but with counts from a single mutation type. A similar procedure can be used to modify LDA to form ILDA, which is similar to the model described by Shiraishi et al. [1], and to modify the CTM to form the ICTM.

Inference

IMMCTM updates are similar to those for the MMCTM [3], with modifications to allow for the independent feature construction of the mutation.

The factorized mean-field variational Bayesian approximation for the IMMCTM is

$$q(\eta, \phi, Z | \lambda, \nu, \theta, \gamma) = \prod_{m=1}^{M} \prod_{k=1}^{K_m} \prod_{i=1}^{I_m} q(\phi_{m_1}^{m_1} | \gamma_{m_1}^{m_1}) \times D \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{k=1}^{K_m} q(\eta_d^{m_1} | \lambda_d^{m_1}, \nu_d^{m_1}) \times D \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{n=1}^{N_{d}^{m_1}} q(z_{dn}^{m_1} | \theta_{dn}^{m_1})$$

(5)

where

- $\phi_{m_1}^{m_1} \sim \text{Dir}(\gamma_{m_1}^{m_1})$
- $\eta_d^{m_1} \sim \mathcal{N}(\lambda_d^{m_1}, \nu_d^{m_1})$
\( z_{dn}^m \sim \text{Cat}(\theta_{dn}^m) \)

The update for \( \gamma_{kij}^m \) is

\[
\gamma_{kij}^m = \alpha_i^m + \sum_{d=1}^{D} \sum_{n=1}^{N_d^m} \theta_{dnk}^m \mathbb{I}(x_{dni} = j)
\]  

(6)

And the update for \( \theta_{dnk}^m \) is

\[
\theta_{dnk}^m \propto \exp \left( \lambda_{dk}^m + \sum_{i=1}^{I} \mathbb{E}_q [\log \phi_{kix_{dni}}^m] \right)
\]  

(7)

where the calculation of \( \mathbb{E}_q [\log \phi_{kix_{dni}}^m] \) is as described in Blei et al. 5

References


