Excess Correlation Analysis: A Spectral Method for Topic Modeling

UBC MLRG Fall 2020

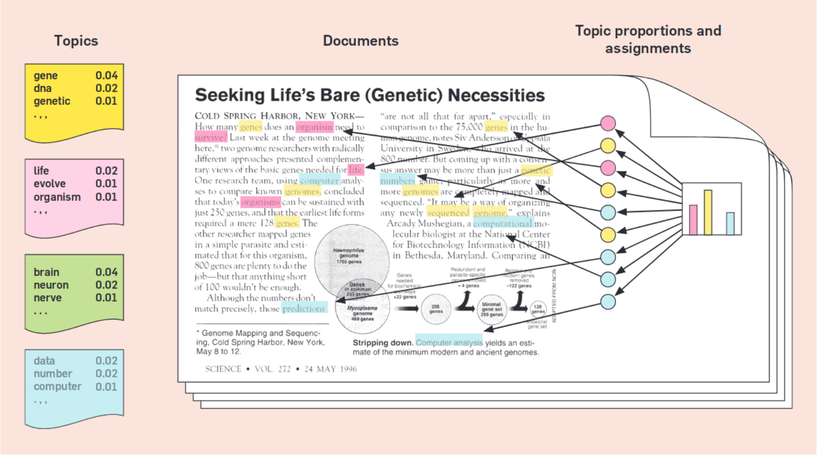
November 4, 2020

Motivation

Overall task: infer a distribution of *topics* based on the contents of a *documents*.

- Given document contents $x_1, x_2, \ldots, x_n$ infer the distribution of topics h.

Figure: The Dirichlet model



Dependencies (from Mark's 540 slides):

$$p(\underbrace{Z}_{\text{topics}}, \underbrace{\pi}_{\text{word prob.}}, \overbrace{\theta}^{\text{topic prop.}} \mid \underbrace{X}_{\text{data (words)}}, \underbrace{\alpha}_{\text{prior on topic proportions}}, \underbrace{\beta}_{\text{prior on word probabilities}}) = \left[\prod_{i=1}^{n} \underbrace{p(\theta^i \mid \alpha)}_{\text{topic proportion probability (document 'i')}} \prod_{j=1}^{d} \underbrace{p(z^i_j \mid \theta^i)}_{\text{topic probability (topic at position 'j' in document 'i')}} \underbrace{p(x^i_j \mid z^i_j, \pi_j)}_{\text{word probability (word at position 'j' in document 'i')}}\right]\left[\prod_{c=1}^{k} \underbrace{p(\pi_c \mid \beta)}_{\text{word probability parameters (topic 'c')}}\right]$$

Traditionally these parameters are learned by Monte Carlo methods like Gibbs sampling. Problem: this could be inefficient.

The Dirichlet model (another perspective) [Anandkumar et. al.]

- $h = (h_1, h_2, \ldots, h_k) \in \mathbb{R}^k$: proportions over topics
- $x_1, x_2, x_3, \ldots \in \mathbb{R}^d$: observed variables (intuition: x_v is what the vth word in the document is)
- matrix $O \in \mathbb{R}^{d \times k}$: usually unknown, so we assume this exists such that $\mathbb{E}(x_v | h) = Oh$ for each $v \in \{1, 2, 3, \ldots\}$

Dirichlet model: h itself follows a Dirichlet distribution over the k topics. Distribution parameterized by $\alpha \in \mathbb{R}^k_{>0}$ as

$$p_\alpha(h) = \frac{1}{Z(\alpha)} \prod_{i=1}^{k} h_i^{\alpha_i - 1}$$

where $Z(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$ is a normalizing constant. As $\sum_i \alpha_i \to 0$ the distribution degenerates so $p_\alpha(h) = 1$ for exactly one h and 0 for the rest.

What are the implications of the O matrix?

LDA assumption: x_v variables take on discrete values out of d outcomes. Intuition here is that the distribution over these d outcomes is dependent on the choice of topic h.
The ith column of O, O_i, is a probability vector of the conditional probabilities of each word under each topic h_i.

$$\mathbb{E}(x_v|h) = \sum_{i=1}^{k} P(t=i|h)\mathbb{E}(x_v|t=i,h) = \sum_{i=1}^{k} h_i \cdot O_i = Oh$$

O is assumed to have full column rank.

Tensors

Traditional methods for LDA involve sampling-based approaches, this paper introduces a method that's based on tensors.
The algorithm hinges on calculating the matrix of second moments and tensor of third moments:

$$\text{Pairs} \in \mathbb{R}^{d \times d} = \mathbb{E}[(x_i - \mu) \otimes (x_j - \mu)]$$

$$\text{Triples} \in \mathbb{R}^{d \times d \times d} = \mathbb{E}[(x_i - \mu) \otimes (x_j - \mu) \otimes (x_l - \mu)]$$

Necessary identities

"Lemma 3.1":

$$\text{Pairs} = \sum_{i=1}^{k} \sigma_i^2 O_i \otimes O_i$$

$$\text{Triples} = \sum_{i=1}^{k} \mu_{i,3} O_i \otimes O_i \otimes O_i \quad \text{where} \quad \mu_{i,3} = \mathbb{E}[(h_i - \mathbb{E}[h_i])^3]$$

Proof for Pairs:

$$\begin{aligned}
\mathbb{E}[(x_1 - \mu) \otimes (x_2 - \mu)] &= \mathbb{E}[\mathbb{E}[(x_1 - \mu|h)] \otimes \mathbb{E}[(x_2 - \mu)|h]] \\
&= O\mathbb{E}[(h - \mathbb{E}[h]) \otimes (h - \mathbb{E}[h])]O^T \\
&= O\mathrm{diag}(\ldots \sigma_i^2 \ldots)O^T
\end{aligned}$$

Excess Correlation Analysis (ECA)

The goal of the algorithm is to estimate the O matrix.

```
procedure ECA(vector θ ∈ ℝ^k, samples x)
    Calculate Pairs = 𝔼[(x_i − μ) ⊗ (x_j − μ)].
    Calculate Triples = 𝔼[(x_i − μ) ⊗ (x_j − μ) ⊗ (x_l − μ)].
    Find a matrix U ∈ ℝ^{d×k} such that range(U) = range(Pairs).
    Find V ∈ ℝ^{k×k} such that V^T(U^T Pairs U)V = I_k.
    Set W ← UV.
    Calculate the left singular vectors Ξ of the matrix W^T Triples(Wθ)W, where Triples(η) = 𝔼[(x_i − μ)(x_j − μ)^T ⟨η, x_l − μ⟩]
    Return the set Ô = {(W^+)^T ξ : ξ ∈ Ξ}.
end procedure
```

Here W^+ denotes the Moore-Penrose inverse of W.

Intuition for ECA

- ECA essentially performs two SVDs.
  - the first (finding W)spherizes the data; the matrix W is supposed to represent data that is "projected" so that it has covariance equal to the identity.
  - the second (explicitly taking the singular vectors of $W^T \text{Triples}(W\theta)W$) is on the third-order moments.
- Taking these SVDs are efficient because the algo only performs the decompositions on $k \times k$ matrices
- The paper's introduction states that the overall purpose of the SVD of the higher-order moment is to find "directions which exhibit non-Gaussianity". It's actually supposed to work for any latent distribution with independent latent factors.

Identities about ECA

Theorem 3.1: Under the independent latent factor model,

- For all $\theta \in \mathbb{R}^k$, the algorithm returns a subset of the columns of O.
- Let $\gamma_i = \mu_{i,3}/\sigma_i^3$ (recall $\mu_{i,3} = \mathbb{E}[(h_i - \mathbb{E}[h_i])^3]$) , and assume $\gamma_i \neq 0$ for each $i \in [k]$. If $\theta \in \mathbb{R}^k$ is drawn uniformly at random from the unit sphere $\mathcal{S}^{k-1}$, then with probability 1, the algorithm returns all the columns of O in canonical form up to sign.

Proof sketch

- This theorem relies on the fact that it is feasible to find the matrix V, which is true because $U^T \mathrm{Pairs} U$ can be shown to be a full-rank matrix. Furthermore the matrix $M = W^T O$ is orthogonal.
- The matrix $W^T \mathrm{Triples}(W\theta) W = MDM^T$ where $D = \mathrm{diag}(M^T \theta)\mathrm{diag}(\gamma_1, \ldots, \gamma_k)$.
- As M is orthogonal we are thus able to find the eigenvalues of this construction, and each singular vector ξ is in the form $s_i M e_i = s_i W^T O_i$. Thus $(W^+)^T \xi = s_i O_i$.

Additional constraints and modifications for LDA

Under the Dirichlet model h has Dirichlet density is indeed a product density, but the h_i are not independent becasue h is constrained so that $\sum h_i = 1$. If we assume $\alpha_0 = \sum_i \alpha_i$ is known then we can define the moments as follows:

$$\mu = \mathbb{E}[x_i]; \mathrm{Pairs}_{\alpha_0} = \mathbb{E}[x_i x_j^T] - \frac{\alpha_0}{\alpha_0 + 1}\mu\mu^T$$

And a modified third moment as

$$\begin{aligned}\mathrm{Triples}_{\alpha_0}(\eta) = \mathbb{E}[x_i x_j^T \langle \eta, x_\ell \rangle] - \frac{\alpha_0}{\alpha_0 + 2}(\mathbb{E}[x_i x_j^T]\eta\mu^T + \mu\eta^T\mathbb{E}[x_i x_j^T] \\ + \langle \eta, \mu \rangle \mathbb{E}[x_i x_j^T]) + \frac{2\alpha_0^2}{(\alpha_0 + 2)(\alpha_0 + 1)}\langle \eta, \mu \rangle \mu\mu^T\end{aligned}$$

Modified ECA algorithm for LDA

procedure ECA(vector $\theta \in \mathbb{R}^k$, samples x)
    Calculate Pairs$_{\alpha_0}$ and Triples$_{\alpha_)}$.
    Proceed as in the original ECA algorithm to find the matrix W and singular values Ξ.
    Return the set

$$\hat{O} = \left\{ \frac{(W^+)^T \xi}{\vec{1}^T (W^+)^T \xi} \mid \xi \in \Xi \right\}$$

end procedure

Theorem 3.2:

- For all $\theta \in \mathbb{R}^k$ using the modified algorithm returns a subset of the columns of O
- If $\theta \in \mathbb{R}^k$ is drawn uniformly at random from the unit sphere $\mathcal{S}^{k-1}$ then the algorithm returns all columns of O with probability 1.
- The Dirichlet parameter α satisfies $\alpha = \alpha_0(\alpha_0 + 1)\mathrm{Pairs}_{\alpha_0}(O^+)^T \vec{1}$ where $\alpha_0 = \sum_i \alpha_i$.

Complexity

Key idea: You need to take enough samples of words in order to come up with an accurate recovery of O. This is the value of "d" and determines the size of Pairs and Triples.

Precisely: If you take

$N \geq C_1 f(\alpha, \sigma_k(O)) = C_1((\alpha_0 + 1)/(\min_i \alpha_i/\alpha_0)\sigma_k(O)^2)$ samples to form empirical versions of the Pairs and Triples constructs, then the algorithm returns a set of columns $\hat{O}_i$ such that

$$\|O_i - \hat{O}_i\| \leq C_2 \frac{(\alpha_0 + 1)^2 k^3}{(\min_i \alpha_i/\alpha_0)\, \sigma_k(O)^3 \sqrt{N}}$$

Here $\sigma_k(O)$ is the kth (minimum) singular value of O.

Discussion

- The algorithm can be seen as a method for obtaining a particular desired decomposition of the tensor Triples, which is $\sum_{i=1}^{k} \mu_{i,3} O_i \otimes O_i \otimes O_i$.
- Paper says that the method in practice is not stable, due to the use of internal randomization.
- Authors suggest that other decomposition methods can be used like "simultaneous diagonalizations of matrices or direct tensor decomposition methods".