CorEx Topic Word Shifts

Ryan J. Gallagher

February 2018

Recall, the overall total correlation TC is the sum over all m topic total correlations,

$$TC = \sum_{j=1}^{m} TC(X_{G_j}; Y_j),$$

and each topic total correlation can be written as the average across all d point-wise total correlations,

$$TC(X_{G_j}; Y_j) = \frac{1}{d} \sum_{k=1}^{d} TC(X = x_k; Y_j).$$

Furthermore, the point-wise total correlation for document k and topic j can be decomposed as

$$TC(X = x_k; Y_j) = \mathbb{E}[\log p(y_j)] + \sum_{i=1}^{n} \alpha_{i,j} \log \frac{p(x_i^{(k)} \mid y_j)}{p(x_i^{(k)})}$$

All together then,

$$\begin{aligned} TC &= \sum_{j=1}^{m} \frac{1}{d} \sum_{k=1}^{d} \left[\mathbb{E}[\log p(y_j)] + \sum_{i=1}^{n} \alpha_{i,j} \log \frac{p(x_i^{(k)} \mid y_j)}{p(x_i^{(k)})} \right] \\ &= \sum_{j=1}^{m} \mathbb{E}[\log p(y_j)] + \sum_{j=1}^{m} \frac{1}{d} \sum_{k=1}^{d} \sum_{i=1}^{n} \alpha_{i,j} \log \frac{p(x_i^{(k)} \mid y_j)}{p(x_i^{(k)})}. \end{aligned}$$

Now, consider when we have a trained CorEx topic model and we have two out-of-sample document-term matrices X_1 and X_2 with shapes $d_1 \times n$

and $d_2 \times n$, respectively. Using the trained topic model, we can estimate TC_1 and TC_2. We are interested in how TC_1 and TC_2 may differ, based on how words additively contributed to each measure of topical information. Specifically, we consider their difference

$$TC_2 - TC_1 = \sum_{j=1}^{m} \frac{1}{d_2} \sum_{k_2=1}^{d_2} \sum_{i=1}^{n} \alpha_{i,j} \log \frac{p(x_i^{(k_2)} \mid y_j)}{p(x_i^{(k_2)})} - \sum_{j=1}^{m} \frac{1}{d_1} \sum_{k_1=1}^{d_1} \sum_{i=1}^{n} \alpha_{i,j} \log \frac{p(x_i^{(k_1)} \mid y_j)}{p(x_i^{(k_1)})}$$

Importantly, we can commute the sum over words to be the outermost sum, yielding

$$TC_2 - TC_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\underbrace{\frac{1}{d_2} \sum_{k_2=1}^{d_2} \alpha_{i,j} \log \frac{p(x_i^{(k_2)} \mid y_j)}{p(x_i^{(k_2)})} - \frac{1}{d_1} \sum_{k_1=1}^{d_1} \alpha_{i,j} \log \frac{p(x_i^{(k_1)} \mid y_j)}{p(x_i^{(k_1)})}}_{\text{Contribution of word } i \text{ to difference in topic } j} \right].$$

Note,

1. We can rank the words in terms of their absolute magnitude in difference, providing us with an interpretable look into how two sets of documents differ in terms of their topical structure.

2. The sign of the contribution indicates whether the topical content of the word is more prevalent in the first or second set of documents.

3. The contributions can be considered across all topics (by summing across all topics), or within a single topic (commuting the sum over topics to be the outermost sum).

4. The contributions can be normalized to fall within the range $[0, 1]$ by dividing by $TC_2 - TC_1$ (or $TC_{2,j} - TC_{1,j}$ when considering contributions within a single topic).

We call the corresponding normalized ranked, signed list of contributions the *topic word shift*.