Investigating Gene Expression Using Correlated Topic Models and Latent Dirichlet Allocation

Ankit Gupta, Harvard University

Motivation

- Textual data and RNA-seq data have surprisingly similar representations
- Latent topics have biological interpretations
- Being able to model biological phenomena through topic modeling allows for useful dimensionality reduction and direct biological interpretation of topics

High Level Process

1. Acquire RNA-seq data as Gene-Counts

Sample	Cel22	Gpx3	Nrsn1	Fbx113	…
Sample1	4	3	2	1	…
Sample3	2	1	0	5	…

2. Apply Topic Model (LDA or CTM)

- Note that gene-count format is very similar to word-counts commonly used in text processing problems
- Use LDA and CTM document-topic distributions as dimensionality reduction
- Determine top genes associated with each topic
- Investigate correlations between topics, if using CTM
- Determine whether correlations correspond to biological phenomena

3. Use new features for classification

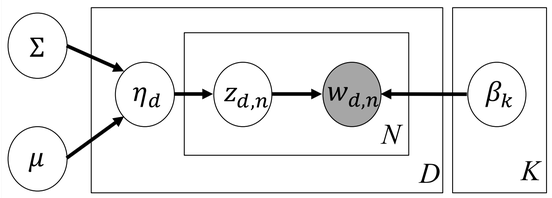
- Use LDA or CTM for dimensionality reduction
- Fit a classifier to low-dimensional data (Logistic Regression, SVM, etc.)
- Compare classification accuracy to the same, using PCA-based features and original features

Key Contributions

- Implemented the inference algorithm for a correlated topic model
- Wrote multicore implementation to parallelize variational inference
- Analyzed classification accuracy using several biological datasets with differing degrees of variability in samples
- Compared results for generated topics and documentation classification against LDA and PCA (implemented elsewhere)

Correlated Topic Model Overview

Graphical Model



Document Generative Process

1) Draw $\eta_d \sim Normal(\mu, \Sigma)$
2) $\theta_d = f(\eta_d)$, to get Logistic Normal
3) for word $n \in \{1, \dots, N_d\}$:
   1. Get topic $Z_{d,n} \sim Mult(\theta_d)$
   2. Get word $w_{d,n} \sim Mult(\beta_{z_{d,n}})$

Goal: Learn μ, Σ, β using variational EM

M-step: Update Model Parameters

$$\hat{\beta}_i \propto \sum_d \phi_{d,i} n_d$$

$$\hat{\mu} = \frac{1}{D}\sum_d \lambda_d$$

$$\hat{\Sigma} = \frac{1}{D}\sum_d diag(\nu_d^2) + (\lambda_d - \hat{\mu})(\lambda_d - \hat{\mu})^T$$

E-step: Variational Inference on Each Document

$$q(\eta_{1:K}, z_{1:N} \mid \lambda_{1:K}, \nu^2_{1:K}, \phi_{1:N}) = \prod_{i=1}^{K} q(\eta_i|\lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n|\phi_n)$$

Using Jensen's inequality, we can bound the log probability of a document as

$$\log p(w_{1:N}|\mu, \Sigma, \beta)$$
$$\geq E_q[\log p(\eta|\mu, \Sigma)] + \sum_{n=1}^{N} E_q[\log p(z_n|\eta)] + \sum_{n=1}^{N} E_q[\log p(w_n|z_n, \beta)\,] + H(q)$$

Update Equations

$$\hat{\zeta} = \sum_{i=1}^{K} \exp(\lambda_i + \frac{1}{2}\nu_i^2)$$

$$\hat{\phi}_{n,i} \propto \exp(\lambda_i)\, \beta_{i,w_n}$$

where ζ was a new variational parameter introduced to calculate gradients.
For the others, we will need to use a gradient-based minimizer, where the gradients are

$$\frac{dL}{d\lambda} = -\Sigma^{-1}(\lambda - \mu) + \sum_{n=1}^{N} \phi_{n,1:K} - \left(\frac{N}{\zeta}\right) \exp(\lambda + \frac{1}{2}\nu^2)$$

$$\frac{dL}{d\nu_i^2} = -\frac{1}{2}\Sigma_{ii}^{-1} - \frac{N}{2\zeta} \exp\left(\lambda_i + \frac{1}{2}\nu_i^2\right) + \frac{1}{2\nu_i^2}$$

Since this process occurs separately on each document in an iteration of the E-step, this was parallelized and run on 64-core machines.

Results

Experiment 1: Proof of Concept: 40 males, 30 females, sex-linked genes
LDA: Produces Y-chromosome topic, and X-chromosome topic
CTM: Produces several Y-chromosome topics and several X-chromosome topics, with negative correlation between X and Y chromosome topics

Experiment 2: Rat Gene Relationships: 320 rat samples, various organs

Method	Organ Prediction Accuracy
Original Features	.98
CTM Features	.93
LDA Features	.875
PCA Features	.89

Interesting topics:
- "Blood-related" genes
- "Heart-related" genes
- "Kidney-related" genes
- "Lipid-related" genes
- "Sugar-related" genes

- Positive covariance between kidney and blood related genes, and negative between kidney and heart
- Need to investigate if it can perform this well on more closely related samples (see Experiment 3)

Experiment 3: Identifying Mouse Neurons Subtypes

Method	Subtype Prediction Accuracy
Original Features	.93
CTM Features	.69
LDA Features	.83
PCA Features	.89

- In differentiating between subtypes, CTM performs poorly
- No clear topics found
- Future experiments will allow for more iterations

Key Takeaways

1) CTM performs well at classification tasks when classes are fairly highly varied
2) CTM produces topics and correlations that appear to correspond with biological phenomena
3) CTM performs poorly relative to LDA when classes are relatively similar, such as subtypes of neurons
4) Slow speed of CTM presents drawback relative to LDA