Bayesian Clustering and Topic Modeling of Gene Expression Data

Skanda Koppula (skoppula@mit.edu), Karren Yang (karren@mit.edu)

6.882 Project Proposal, April 17, 2017

Overview

We explored the usage of various topic and clustering models in the analysis of gene expression data. For the former, we explored whether topic modeling could identify biologically relevant 'topics'. In this context, a 'topic' is a set of genes that together perform a specific biological function (a 'gene module'); we compared our results to modules found by biology literate. For the latter, we explored whether we are able to cluster cells accurately based on their gene expression levels. [1].

Topic Model: Latent Dirichlet Allocation

We were unable to yield any results using Python's out-of-the-box implementation of `lda` [3]. With our 250 MB dataset, the collapsed Gibbs sampler used in the implementation was taking too long to produce samples, even on larger server-grade machines and with a small number of LDA clusters. We explored modifying the source to parallelize the sampling, but found the source to be crabbed and hard to follow.

In our search for an alternative, we explored another implementation that used online variational bayes for posterior estimate [4, 6], and a broken `C++` Gibbs sampler for LDA that we modified to work [5]. We were able to get this latter sampler running parallel across multiple cores, with a burn-in of 100 iterations. We compare these two posterior estimation approaches using our entire dataset, with a 10% held-out testing partition. We experimented with $k = 5, 10, 25,$ and 50 clusters.
Figure **??** in the Appendix shows the time to complete each estimation method. As expected, Gibbs scales poorly with the parameter dimensionality and is strictly worse than online variational Bayes across all studied topic counts.
Figure **??** in the Appendix shows the time to complete each estimation method. As expected, Gibbs scales poorly with the parameter dimensionality and is strictly worse than online variational Bayes across all studied topic counts.
Note that these p-values do *not* factor for multiple hypothesis corrections, so could be misleadingly significant.

Division of Work

Remaining Work and Schedule

- robust to dropout - time series

Appendix

References

[1] The XL-mHG Test For Enrichment: A Technical Report. `https://arxiv.org/pdf/1507.07905.pdf`

[1] The genes in gene modules generally move in tandem: so when a gene module is upregulated, all genes have higher expression (i.e. frequency). This motivates our use of topic models to capture this relationship.

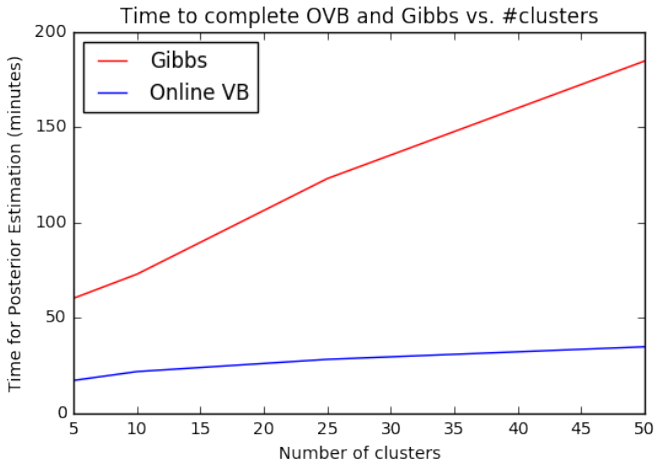

Figure 1: Comparison of the running times of each of the posterior estimation methods across various cluster sizes.

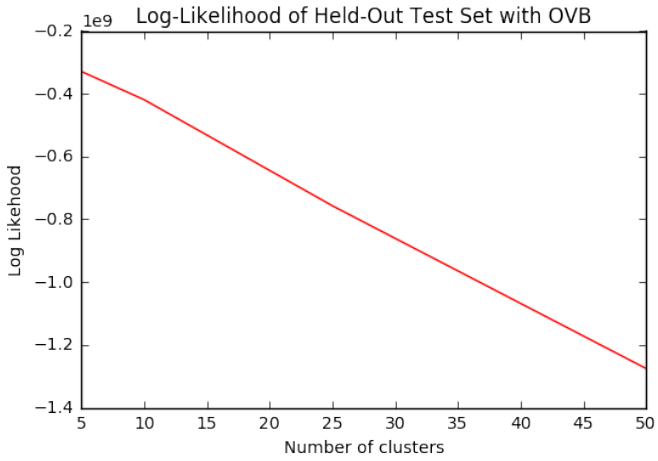

Figure 2: Log-likelihood of the held-out testing set, across various cluster sizes.

[2] Molecular Signatures Database v6.0. http://software.broadinstitute.org/gsea/msigdb

[3] lda: Topic modeling with latent Dirichlet Allocation. http://pythonhosted.org/lda/

[4] Online Latent Dirichlet Allocation with variational inference. https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/decomposition/online_lda.py

[5] C++ implementation of Latent Dirichlet Allocation https://github.com/openbigdatagroup/plda/blob/master/lda.cc

[6] Online Learning for Latent Dirichlet Allocation. https://pdfs.semanticscholar.org/157a/ef34d39c85d6576028f29df1ea4c6480a979.pdf

[7] Hierarchical Dirichlet Processes http://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf

[8] The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling http://www.cs.columbia.edu/~blei/papers/WilliamsonWangHellerBlei2010.pdf

[9] A Time-Series DDP for Functional Proteomics Profiles https://www.ma.utexas.edu/users/pmueller/pap/NM12.pdf