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 mean-field 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 1 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.

We had time to calculate perplexity and log-likelihood for the OVB parameters, and very surprisingly, we find an increasing log-likelihood for larger cluster sizes on a held-out test set (Figure 2). We are currently trying to understand whether this is because of programming error or a legitimate result.

As is standard in computational gene module method validation, we use a hypergeometric test to compare our topics' collections of genes, with existing collections of genes catalogued in the comprehensive gene module database, `MSigDB` [2]. Figure 3 in the Appendix shows topic-to-`MSigDB` module matches for which the p-value of the match is at least less than 0.3. Models with larger clusters tend to have more matches with higher significance. There a very slightly higher number of matches using the topics discovered by parallelized Gibbs. There are repeat matches, which is slightly disheartening. We note however that these p-values do *not* factor for multiple hypothesis corrections (yet), so could be misleadingly significant!

Clustering Model: Finite-Sized Mixture Model

We are currently implementing from scratch the mixture model specified by the plate model in Figure 4. The underlying assumption of behind the model is that the cell cluster assignment, z_i, would be indicative

[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.

of it's gene expression values as well. Our implementation uses `numpy`, and we are exploring backends to make posterior inference computationally feasible (our current implementation, which uses a Gibbs sampler, overflows memory, so we are implementing stochastic variational methods, like used in our LDA experiments).

Remaining Work and Schedule

We are making progress as per the schedule outlined in our proposal. Apart from continuing with work on the finite-sized mixture model, we plan on running a few experiments to test robustness of our methods to dropout, and scaling IBP and HDP to the full dataset.

Appendix

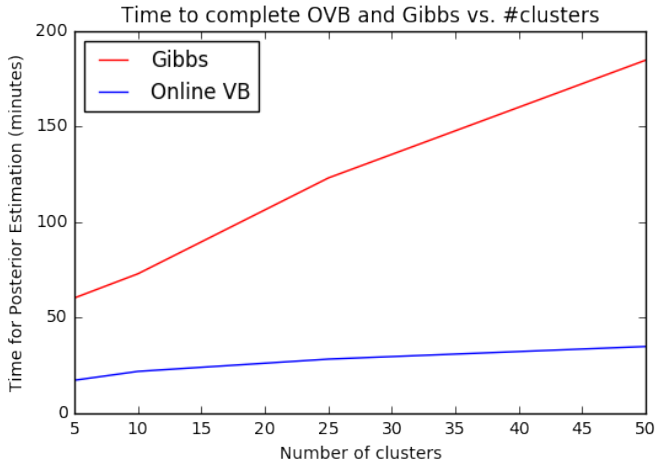

Figure 1: Comparison of the running times of each of the posterior estimation methods across various cluster sizes.

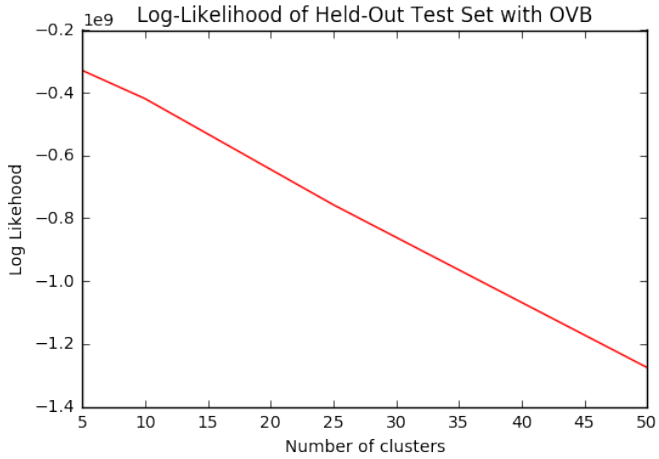

Figure 2: Log-likelihood of the held-out testing set, across various cluster sizes.

```
GIBBS TOPICS
        cluster_size: 5
        cluster_size: 10
                match: topic  1 ,  KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_UP , pval: 0.041958041958
                match: topic  1 ,  PILON_KLF1_TARGETS_DN , pval: 0.137062937063
                match: topic  6 ,  GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN , pval: 0.263736263736
        cluster_size: 25
                match: topic  8 ,  PILON_KLF1_TARGETS_DN , pval: 0.263736263736
                match: topic  22 ,  KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_UP , pval: 0.041958041958
                match: topic  22 ,  PILON_KLF1_TARGETS_DN , pval: 0.137062937063
                match: topic  22 ,  GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN , pval: 0.263736263736
        cluster_size: 50
                match: topic  2 ,  ROME_INSULIN_TARGETS_IN_MUSCLE_UP , pval: 0.193006993007
                match: topic  3 ,  PILON_KLF1_TARGETS_DN , pval: 0.00699300699301
                match: topic  6 ,  PILON_KLF1_TARGETS_DN , pval: 0.263736263736
                match: topic  7 ,  BLALOCK_ALZHEIMERS_DISEASE_UP , pval: 0.018648018648
                match: topic  17 ,  KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_UP , pval: 0.041958041958
                match: topic  17 ,  PILON_KLF1_TARGETS_DN , pval: 0.193006993007
                match: topic  22 ,  PILON_KLF1_TARGETS_DN , pval: 0.153846153846
                match: topic  24 ,  WEI_MYCN_TARGETS_WITH_E_BOX , pval: 0.0839160839161
OVB TOPICS
        cluster_size: 5
        cluster_size: 10
                match: topic  5 ,  PILON_KLF1_TARGETS_DN , pval: 0.263736263736
        cluster_size: 25
                match: topic  5 ,  DIAZ_CHRONIC_MEYLOGENOUS_LEUKEMIA_UP , pval: 0.263736263736
                match: topic  8 ,  KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_UP , pval: 0.041958041958
                match: topic  15 ,  PUJANA_ATM_PCC_NETWORK , pval: 0.263736263736
                match: topic  15 ,  PUJANA_BRCA1_PCC_NETWORK , pval: 0.263736263736
        cluster_size: 50
                match: topic  6 ,  GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN , pval: 0.263736263736
                match: topic  15 ,  PUJANA_ATM_PCC_NETWORK , pval: 0.263736263736
                match: topic  22 ,  MARSON_BOUND_BY_FOXP3_UNSTIMULATED , pval: 0.263736263736
                match: topic  28 ,  PILON_KLF1_TARGETS_DN , pval: 0.201398601399
                match: topic  38 ,  PUJANA_BRCA1_PCC_NETWORK , pval: 0.263736263736
                match: topic  41 ,  MARSON_BOUND_BY_FOXP3_UNSTIMULATED , pval: 0.263736263736
                match: topic  42 ,  KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_UP , pval: 0.193006993007
```

Figure 3: Matches between the gene collections found in LDA topics and published gene sets in `MSigDB`.

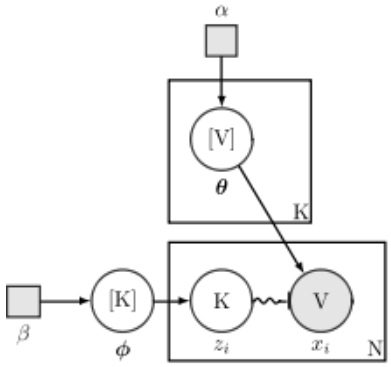

Figure 4: Finite `K`-sized mixture model currently implemented. θ is the parameter for every cluster component, represented from a categorical draw of over all genes. z_i is the cluster assignment, and ϕ is the distribution of clusters. As usual, α, β are hyper-parameters.

References

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

[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