Bayesian Clustering and Topic Modeling of Gene Expression Data

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

6.882 Project Proposal, April 18, 2017

Overview

We have made preliminary assessments of three topic models – LDA, the Hierarchical Dirichlet Process (HDP), and the Indian Buffet Process Compound Dirichlet Process (IDP-DP) – for discovering topics in single-cell RNA-sequencing (scRNA-seq) data. In this context, topics are sets of genes that are co-expressed across cells (i.e. gene modules) [1]. Since we would like topics to correpond to biologically useful gene modules, our main evaluation metric is the correpondence of the learned topics to known functional gene sets curated from literature (i.e. from the `MSigDB` database). While HDP and the IBP-DP offer greater flexibility in that the number of topics does not have to be defined a priori, we do not find evidence that they yield significantly more coherent topics; moreover, we find in practice that training the models and tuning the hyper-parameters to maintain a reasonable number of topics is less time-efficient than applying LDA over a range of number of topics. Moving forward, we will focus on implementing and evaluating a unified framework for clustering and topic modeling based on LDA.

Results

Running LDA We explored two implementations of posterior inference for LDA on large datasets: one using 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 $k = 5, 10, 25, 50$ topics.

Figure 1 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 also calculated perplexity and log-likelihood for the OVB parameters, and very surprisingly, we found that log-likelihood decreased as number of topics increased on a held-out test set (Figure 2). We are currently trying to understand whether this is because of programming error.

Evaluating Biological Significance of LDA Topics We did gene set enrichment analysis using the minimum hypergeometric test [1] to compare our topics, which are ranked lists 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 more topics tended to have more matches with higher significance. These results suggest that we need to train the model with more topics. We note that the p-values do *not* factor for multiple hypothesis corrections, so at the moment we are only using them as relative measures of model quality.

LDA vs. HDP We trained an HDP topic model and compared it to LDA models with similar numbers of topics, using held-out perplexity and gene set enrichment analysis as described above as our evaluation metrics. Since the available implementation of the posterior inference algorithm for HDP was too slow to run on the entire dataset, we did feature selection using the Seurat toolkit in R [11] to reduce the number of genes in our dataset with low variance across cells, and trained all models on this subset of data. We tuned

[1] The genes in gene modules generally move in tandem: so when a gene module is upregulated, all genes have higher expression. This motivates our use of topic models to capture these relationships.

the concentration parameters in HDP to obtain a reasonable number of topics (i.e. 150); subsequently, we trained LDA models with similar numbers of topics for comparison. We found that the HDP model had higher perplexity on a held-out dataset (Figure 4). Moreover, the topics in the HDP model tended to have less significant enrichment for known gene sets from `MSigDB`, and we found that this difference to be significant (Figure 5, Mann-Whitney-U test, $p < 0.0001$). Overall, the LDA model proved superior to the HDP model in this experiment.

LDA vs. IBP-DP One drawback to the HDP is that there tends to be a correlation between how frequently a topic appears across all documents and how prevalent this topic is within documents that it appears in. Williamson et al. [8] proposed the 'focused topic model' to overcome this drawback. In their model, each topic $k = 1, 2, ...$ has a relative prevalence $\phi_k \sim Gamma(\gamma, 1)$ and a population frequency $\pi_k = \prod_{j=1}^{k} \mu_k$, where each $\mu_k \sim Beta(\alpha, 1)$. For each document m, whether topic k appears is sampled as $b_{mk} \sim Bernoulli(\pi_k)$, and the topic proportions are sampled as $\theta_m \sim Dirichlet(b_m \cdot \phi)$.

Since the code from the original IBP-DP paper was not available, we implemented an inference algorithm using collapsed Gibbs sampling [8]. Due to the non-conjugacy of the model, sampling each latent variable from its full conditional required using another sampling method. To sample the topics parameters π and ϕ, we used slice sampling based on the semi-ordered stick-breaking representation of the model [10].

We tested our code on a subset of the Reuters-21578 dataset, using several different values of the concentration hyper-parameter α, which influences the number of clusters. Although higher values of α yielded better log-likelihood values, we found that it resulted in a large number of very small topics, which are not very useful (Figure 6). Qualitatively, we did not find the topics from IBP-DP (Figure 7) to be more coherent than topics than LDA (Figure 8). The most prevalent topics from the IBP-DP each corresponded to similar topics from LDA; less prevalent topics tended to consist of a few unrelated words. These results discouraged us from optimizing the code to train this model on our scRNA-seq dataset, as we do not think it would yield more coherent gene modules than LDA. We emphasize that these results are not completely unexpected, as the authors of the IBP-DP paper did not show any topics from their model, nor did they assess the quality of their model with metrics other than perplexity.

Remaining Work and Schedule

In the next month, we will focus on implementing and evaluating a unified framework for clustering and topic modeling based on LDA. Currently, non-negative matrix factorization appears to be the primary unsupervised method for combined learning of gene modules and cell clusters from scRNA-seq data (ref); typically, the pipeline for scRNA-seq analysis involves first clustering cells, and then using differential analysis to find gene modules that distinguish each cluster (ref). We are currently exploring and implementing Bayesian mixture models for clustering cells; specifically, we are implementing the finite mixture model specified by the plate diagram in Figure 9. Our implementation uses `numpy`, and we are exploring backends to make posterior inference computationally feasible. If we are successful, we will combine these models with LDA and test our unified framework against existing methods of cell clustering and gene module discovery.

Appendix

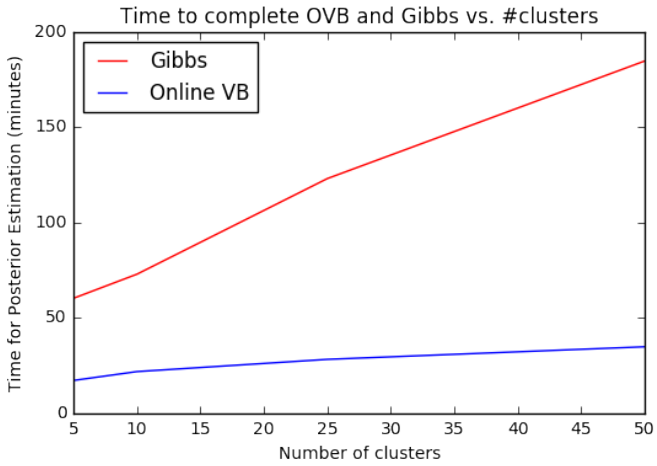

Figure 1: Comparison of the running times of each of the posterior estimation methods across various numbers of topics.

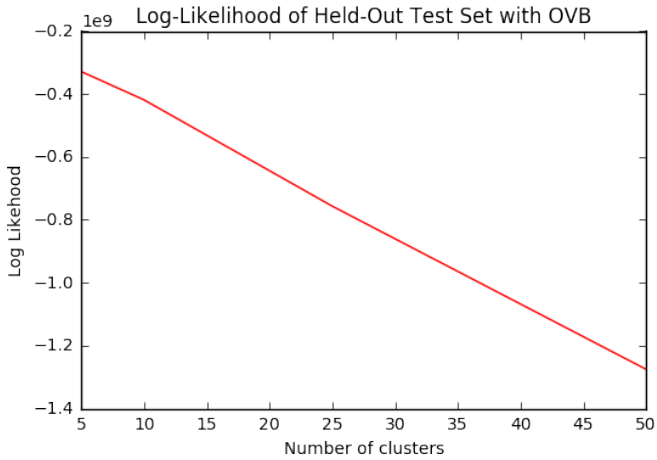

Figure 2: Log-likelihood of the held-out testing set, across various numbers of topics.

```
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`. Cluster size refers to the number of topics in the model

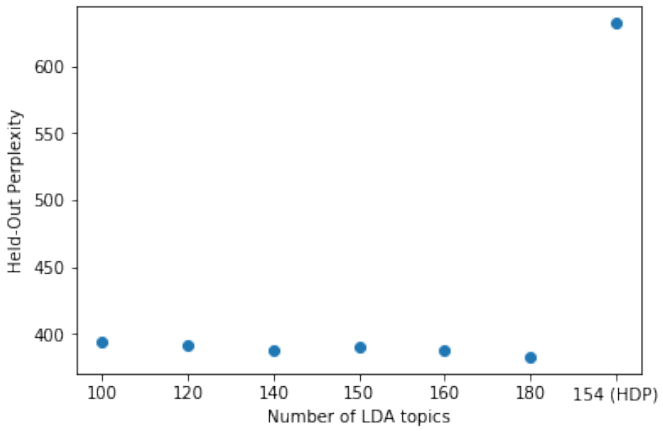

Figure 4: Comparison of log-likelihood of the held-out testing set, under various LDA models and the HDP model.

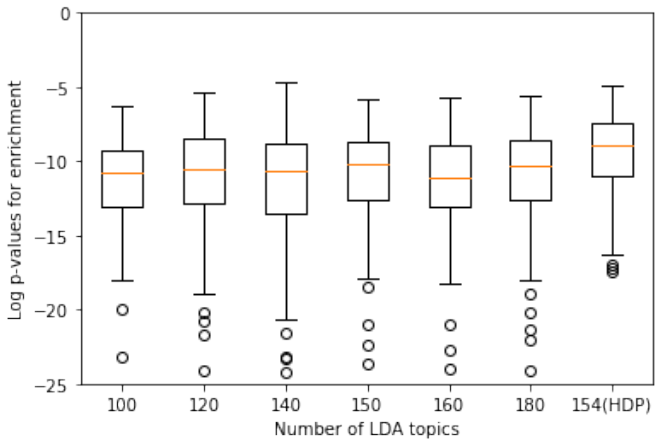

Figure 5: Comparison of distributions of p-values from gene set enrichment analysis between LDA models and the HDP model.

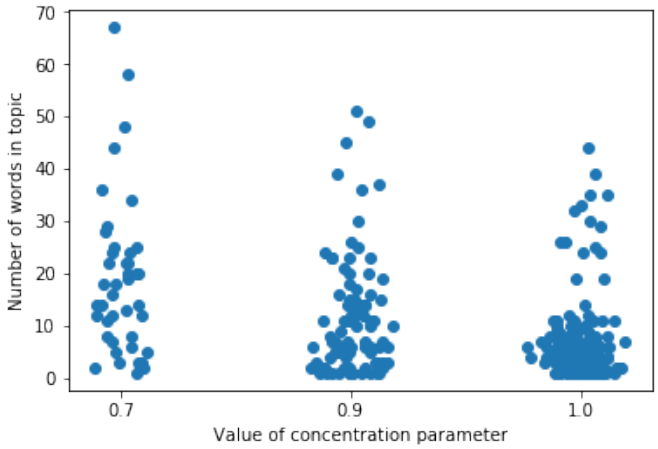

Figure 6: Beeswarm plot of number of words per topic, for 3 different IBP-DP models with different concentration parameters. Each point represents one topic from its model

```
Topic 0: charles diana prince royal parker bowles camilla queen family marriage public princess love britain england
Topic 1: church told during year time world very years made life saying last became take while
Topic 2: catholic n't bishop son women father love mother television day cardinal britain woman leaders days
Topic 3: sunday including against day won group held known last police official roman few come late
Topic 4: pope france visit john french paul first both pontiff trip church religious catholics king against
Topic 5: mother teresa doctors home heart charity hospital order tuesday people work told peace world house
Topic 6: set years ceremony led german germany second made called around rights married time few south
Topic 7: pope health vatican mass trip reporters surgery during saturday death left past spokesman people monday
Topic 8: people president state last later took around government good since say percent age won head
Topic 9: media later catholic former own head newspaper next known leader against wednesday taken n't called
Topic 10: official added first health ago under among around wednesday paul few minister monday children mother
Topic 11: former end throne century taken reports children newspaper england home french paris 1992 international state
Topic 12: french members visit leaders family germany national statement say three work against thursday year
Topic 13: four heart clinton percent age
Topic 14: life say wednesday left while month tuesday times later official take peace paris doctors including
Topic 15: says 1992 princess love left government political under german married took since saturday house around
Topic 16: bernardin cardinal among death u.s own surgery told minister doctors great several news end until
Topic 17: reports roman us told president times
Topic 18: paul white show reports end son union month monday bishop long
Topic 19: service end million since wednesday roman home wife spokesman son reports city
Topic 20: first local place princess british peace white saturday taken expected several united married made century
Topic 21: part say sunday members year leaders church days
Topic 22: former ago n't three century rome funeral group year saying led family
Topic 23: france home news south whose president work east take first both united women country officials
Topic 24: four week throne saying became former died members camilla made long show
Topic 25: statement french week thursday war christian born vatican house heart leader britain 1992 set three
Topic 26: minister prime expected group officials union died children times michael whose off around american church
Topic 27: years president
Topic 28: diana funeral service princess reuters hospital son died
Topic 29: monday u.s around several four children throne year since john tuesday members churchill told statement
Topic 30: day friday party private wednesday later british officials former family until throne white
Topic 31: against
Topic 32: city international good prize second won paris take since died house off local years
Topic 33: times children world
Topic 34: paris work
Topic 35: michael king paul local father political show ceremony next part private war german week whose
Topic 36: rights government prize service million become officials held police head reuters political us party
Topic 37: television reporters show n't political times clinton several off own government son wednesday very during
Topic 38: known news
Topic 39: part great mass later women past says
Topic 40: told charles called off born
Topic 41: taken ceremony president
Topic 42: thursday long leaders
```

Figure 7: Topics from IBP-DP model trained on subset of Reuters dataset.

```
Topic 0 : france home news years work mother president during take whose women east country love told
Topic 1 : church years cardinal bishop take england against million vatican past news british told sunday ceremony
Topic 2 : pope health mass during visit saturday trip told john paul pontiff people church service spokesman
Topic 3 : mother teresa heart sunday home hospital tuesday told doctors order people catholic charity peace house
Topic 4 : pope france french visit church trip first paul catholic pontiff both john state including paris
Topic 5 : teresa mother doctors charity official hospital home work first around told world during under saying
Topic 6 : church media michael paul former marriage princess england never n't love very told public years
Topic 7 : bishop church catholic son father n't told women love mother woman roman years leaders ago
Topic 8 : royal family queen prince charles throne church princess century britain first british media 1992 head
Topic 9 : order day city friday group during own doctors monday very reuters prize last people roman
Topic 10 : television told show n't reporters president later day own times political clinton off year years
Topic 11 : president rights government people last church says state life died told political country group catholic
Topic 12 : diana charles princess britain time wednesday ago family monday million camilla church newspaper bowles parker
Topic 13 : charles parker bowles prince camilla diana royal marriage public queen love king church woman family
Topic 14 : pope bernardin vatican church surgery health year time left told life say death cardinal made
```

Figure 8: Topics from LDA model with 15 topics trained on subset of Reuters dataset.

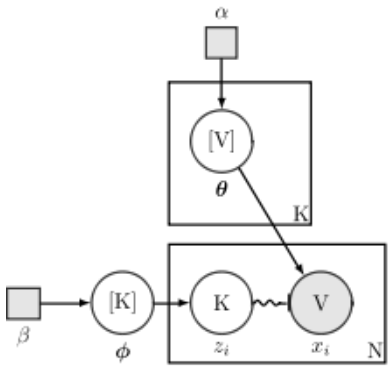

Figure 9: Finite K-sized mixture model currently implemented. θ is the parameter for every cluster component, represented from a categorical draw of over all genes. z_i is the cluster assignment, and ϕ is the distribution of clusters. As usual, α, β are hyper-parameters.

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

[10] Stick-breaking Construction for the Indian Buffet Process http://mlg.eng.cam.ac.uk/zoubin/papers/TehGorGha07.pdf

[11] Spatial reconstruction of single-cell gene expression data http://www.nature.com/nbt/journal/v33/n5/full/nbt.3192.html