Bayesian Clustering and Topic Discovery: Adventures with Gene Expression Data

Karren Dai Yang, Skanda Koppula {KARREN, SKOPPULA}@MIT.EDU
Massachusetts Institute of Technology, Cambridge, MA 02139 USA

1. Introduction

Tumors cell lines are composed of different subpopulations of cells which often exhibit shared patterns of gene expression. Biologists are interested in two key questions: given gene expressions values, (1) can we identify cell clusterings, and (2) can we identify clusters of biologically-related genes? Our goal with this project was to answer both these two questions using Bayesian methods.

In brief, we explored the use of three Bayesian clustering methods (a vanilla mixture, integrated topic-mixture, and non-parametric models) to address the first question, and two topic models (vanilla and a dynamic-topic LDA) to address the second.

1.1. Description of Data

Using a contemporary gene sequencing machine, we obtained samples of single-cell RNA-sequencing data (scRNA-seq) taken from tumors in mice. Our data consisted of the expression values of 22,712 genes for each of 4645 cells. In total, the dataset amounted to 0.86 GB, presenting significant computational challenges when attempting posterior estimation. Apart from a few computational tricks (online inference, multi-core parallelization), for most methods, we pruned non-informative genes from the dataset, ranking based on the deviation in value across cells [1]. The Seurat biological toolkit in R was also used for the purpose of low-variance feature selection (Satija, 2012). We eschewed dimensionality reduction techniques such as PCA because of loss of its direct feature interpretability. Prior researchers have labeled the cells in our dataset; there are a total of 9 cell categories. The complete dataset, as well our preprocessed and pruned versions, are openly available (Yang & Koppula).

[1] We recognize that this can bias towards noisy genes. We favor this method because it is simple and easy to implement, and a practice used in literature (Ling, 2012)

1.2. Prior work

Prior research has explored the use of various computational techniques to analyze gene expression data. Most commonly, Spearman and Pearson correlation metrics are frequently used infer sets of genes that cluster together (Xie, 2015; Borenszstein, 2017). Other techniques, including PCA followed by linear regression, has been used for expression-based cell clustering (Stegle, 2016). Yu et al. propose an unsupervised classifier ensemble as another approach to cell clustering (Yu, 2016).

More Bayesian approaches have also been tried in prior work from the Pe'er lab. (Prabhakaran, 2016) uses a Heirarchical Dirichlet Mixture Model to learn cell clusterings. (Azizi, 2017) builds on this to jointly learn optimal normalization pre-processing of the data. Bayesian networks have also been used in an attempt to learn gene dependencies from expression data (Pe'er, 2000). Our work uses different models to explore gene expression data, but where appropriate (e.g. in mixture models in Section 3), we compare results.

1.3. Structure of Report

We first discuss our experiments using Bayesian topic models to discover related sets of genes in a 'topic': LDA in Section 2 and Dynamic Time Models in Section 4. Then, we discuss our experiments in clustering: Dirichlet Mixtures in Section 3, Integrated Topic-Clustering in Section 5, and non-parametric models in Section 6. We conclude our paper with our observations from across all our studies.

For the purpose of reproducibility, all code can be found at `https://github.com/skoppula/882`.

2. Latent Dirichlet Allocation

2.1. Model Description

In the generative process for LDA, the topic assigned to each word is a drawn from the document's topic

distribution. The identity of the word is drawn from topic's word distribution. Assuming the reader is familiar with LDA, we relegate further details and formalization of the model to (Blei, 2003).

In the context of analyzing gene expression data, we are interested in discovering 'topics' that comprise of a set of top-N genes within the topic distribution that are biologically related. For example, together the genes may direct a specific chemical function in a cell. Biologists denote such sets of genes as 'gene modules' which can be cross-referenced with existing gene module databases.

2.2. Implementation

A first attempt using the built-in Python `lda` package resulting in early memory overflows during what we suspect was pre-allocation of per-document variables. The source code was not available, so we had few clues.

We switched to two open-source implementations: an online mean-field variational Bayes for posterior estimation (Nothman, 2017), and a broken `C++` Gibbs sampler for LDA (OpenDataGroup, 2015).

We fixed portions of the sampler to compile properly and extended the sampler to run across four cores. Details of our sampling procedure can be found in Appendix 7.2. We compared these two posterior estimation approaches using our entire dataset, using a 10% held-out testing partition. We experimented using $k = 5, 10, 25, 50$ topics. We did not require dataset pruning after these optimizations.

2.3. Experiments

In lieue of intrusion testing, we evaluated the interpretability of our topics (ranked lists of genes) using using hypergeometric tests (Wagner, 2015). In brief, this determines the probability that our sets of genes would be clustered together in accordance to with existing collections of genes catalogued by biologists in the gene module database `MSigDB` (Broad-Institute, 2015).

Figure 3 in the Appendix shows gene module matches in `MSigDB` for which the p-value of the match is at least less than 0.3. Notice that models with more topics tended to have more matches with higher significance. The p-values in our tests do *not* factor for multiple hypothesis corrections, so at the moment we are only using them as relative measures of model quality.

It is validating to see that many of the similar modules, such as the `BRCA1` and `DOXORUBICIN`, are cancer related, given that our expression assay was from tumor cells.

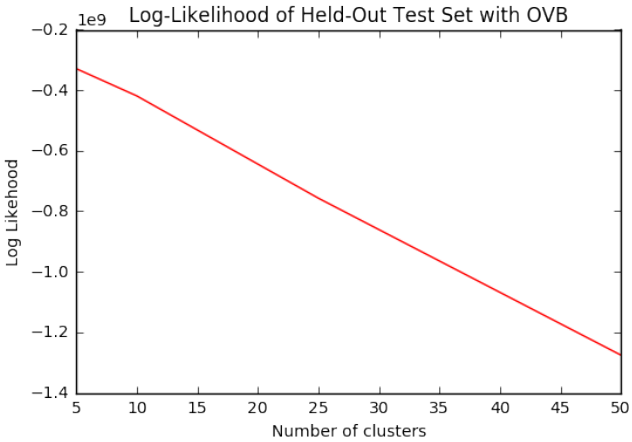

Figure 1. Log-likelihood of the held-out testing set, across various numbers of topics.

Extracting the parameters of each model, we calculated log-likelihood for on a held-out test set. This is shown in Figure 1. We found that for this dataset increasing the topic count increases log-likelihood, roughly logarithmically from 5 to 50. Further experiments will need to test at what point log-likelihood drops off (one way to optimize the number of topics).

As an interesting aside, Figure 2 shows the time until inference completion (collected during our experiments). Gibbs appears to scale poorly with the parameter dimensionality, in contrast to online variational Bayes [2].

Posterior predictive checks to test the mutual information between the each cell and its words' topics is something that we did not have time, but would be interesting to examine. It's not clear that this independence assumption is true in the context of gene expression data, because of cross-talk and regulatory mechanisms between gene modules.

- posterior predictive checks - no intrusion testing meaningful - hypergeometric tests - likelihood - performance metrics

[2] It is important to note that comparing the magnitude of the time to complete inference is not particularly meaningful; as described in Appendix 7.2, we chose a reasonable guess for the number of iterations after which to stop estimation

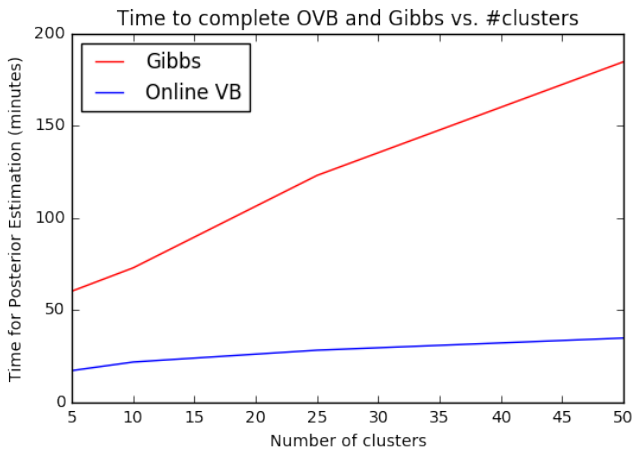

Figure 2. Comparison of the running times of each of the posterior estimation methods across various numbers of topics.

Algorithm 1 Mixture Model

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

3. Dirichlet Mixture Model

3.1. Model Description

Algorithm 1 describes the generative process for the mixture odel.

3.2. Implementation

3.3. Experiments

- comparison with pe'er paper

4. Dynamic Time Model

4.1. Model Description

4.2. Implementation

4.3. Experiments

5. Integrated Topic-Clustering Model

5.1. Model Description

5.2. Implementation

5.3. Experiments

6. Non-parametric Models: IBP and HDP

6.1. Model Description

6.2. Implementation

6.3. Experiments

References

Azizi, et al. Bayesian inference for single-cell clustering and imputing. https://genomicscomputbiol.org/ojs/index.php/GCB/article/view/46/35, Genomics 2017.

Blei, et al. Latent dirichlet allocation. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf, JMLR 2003.

Blei, et al. The IBP compound dirichlet process and its application to focused topic modeling. http://www.cs.columbia.edu/~blei/papers/WilliamsonWangHellerBlei2010.pdf, ICML 2010.

Borenszstein, et al. Xist-dependent imprinted x inactivation and the early developmental consequences of its failure. http://www.nature.com/nsmb/journal/v24/n3/fig_tab/nsmb.3365_SF1.html, Nature Structural & Molecular Biology 2017.

Broad-Institute. Molecular signatures database v6.0. http://software.broadinstitute.org/gsea/msigdb, PeerJ 2015.

Hoffman, et al. Online learning for latent dirichlet allocation.

Ling, et al. Improving relative-entropy pruning using statistical significance. https://www.cs.cmu.edu/~awb/papers/coling2012/rep_coling2012.pdf, ACL 2012.

Nothman, Joel. Online latent dirichlet allocation with variational inference. https://github.

com/scikit-learn/scikit-learn/blob/master/sklearn/decomposition/online_lda.py, 2017.

OpenDataGroup. C++ implementation of latent dirichlet allocation. https://github.com/openbigdatagroup/plda/blob/master/lda.cc, 2015.

Pe'er, et al. Using bayesian networks to analyze expression data. http://www.cs.huji.ac.il/~nir/Papers/FLNP1Full.pdf, Genomics 2000.

Prabhakaran, et al. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. http://www.c2b2.columbia.edu/danapeerlab/html/pub/prabhakaran16.pdf, ICML 2016.

Satija, et al. Spatial reconstruction of single-cell gene expression data. http://www.nature.com/nbt/journal/v33/n5/full/nbt.3192.html, Nature Biotechnology 2012.

Stegle, et al. Computational and analytical challenges in single-cell transcriptomics. https://www.nature.com/nrg/journal/v16/n3/full/nrg3833.html, Nature Methods 2016.

Teh, et al. Stick-breaking construction for the indian buffet process. http://mlg.eng.cam.ac.uk/zoubin/papers/TehGorGha07.pdf, NIPS 2007.

Wagner, Florian. The xl-mhg test for enrichment: A technical report. https://arxiv.org/pdf/1507.07905.pdf, PeerJ 2015.

Xie, et al. SINCERA: A pipeline for single-cell rna-seq profiling analysis. http://months.plos.org/ploscompbiol/article?id=10.1371/month.pcbi.1004575, PLoS 2015.

Yang, Karren and Koppula, Skanda. Tumor cell melanoma data. https://www.dropbox.com/sh/tdrplhiu3mzvmeu/AADFlZITyPX_-I4O7jDuruJra?dl=0.

Yu, et al. SC3 - consensus clustering of single-cell rna-seq data. http://biorxiv.org/content/early/2016/09/02/036558, Nature Methods 2016.

7. Appendix

7.1. Breakdown of Work

The implementation(s) and experiments involving LDA, Dirichlet Mixtures, and Dynamic Time Model was completed by Skanda (Section 2, Section 3, and Section 4). Karren completed the implementations and experiments involved Integrated Topic-Clustering, Non-Parametric models, and the shared gene enrichment test implementation (Section 5 and Section 6. We believe the authors contributed equally in this work.

7.2. Latent Dirichlet Allocation

For our Gibb's sampler, we had a fixed burn-in of number of samples (200), a fixed number of sampling iterations after that (500). We didn't extensively explore varying these values, but trying out significantly more iterations (700) didn't seem to change the topic's word distributions significantly. There was one sampling chain on each of four cores.

7.3. 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. (Blei, 2010) 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 (Blei, 2010). 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 (Teh, 2007).

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

```
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

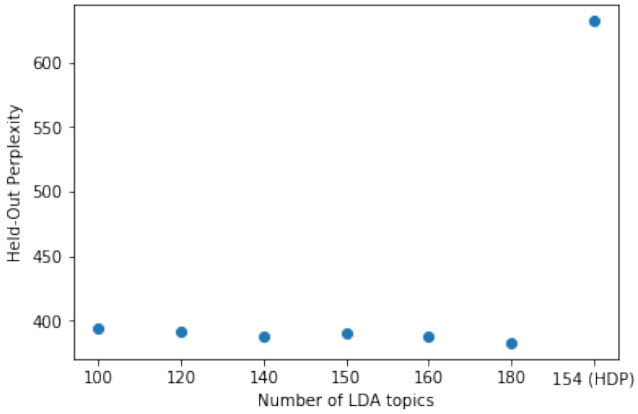

Figure 4. Comparison of log-likelihood of the held-out testing set, under various LDA models and the HDP model.

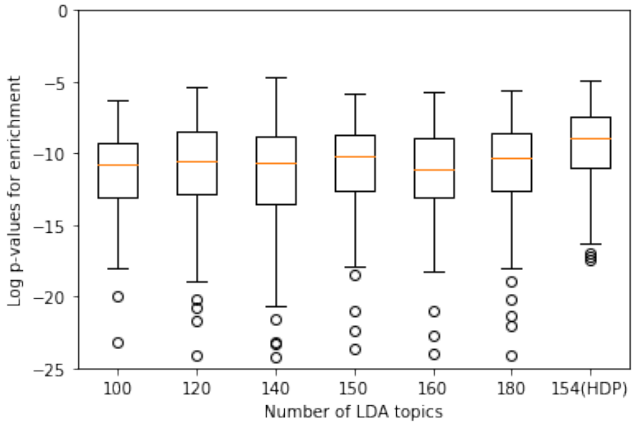

Figure 5. Comparison of distributions of p-values from gene set enrichment analysis between LDA models and the HDP model.

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.

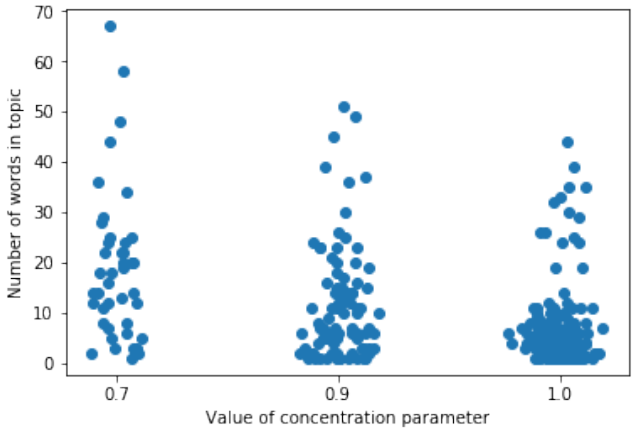

Figure 6. Beeswarm plot of number of words per topic, for 3 different IBP-DP models with different concentration parameters. Each point represents one topic from its model

```
Topic 0: charles diana prince royal parker bowles camilla queen family marr
Topic 1: church told during year time world very years made life saying las
Topic 2: catholic n't bishop son women father love mother television day ca
Topic 3: sunday including against day won group held known last police offi
Topic 4: pope france visit john french paul first both pontiff trip church
Topic 5: mother teresa doctors home heart charity hospital order tuesday pe
Topic 6: set years ceremony led german germany second made called around ri
Topic 7: pope health vatican mass trip reporters surgery during saturday de
Topic 8: people president state last later took around government good sinc
Topic 9: media later catholic former own head newspaper next known leader a
Topic 10: official added first health ago under among around wednesday paul
Topic 11: former end throne century taken reports children newspaper englar
Topic 12: french members visit leaders family germany national statement sa
Topic 13: four heart clinton percent age
Topic 14: life say wednesday left while month tuesday times later official
Topic 15: says 1992 princess love left government political under german ma
Topic 16: bernardin cardinal among death u.s own surgery told minister doct
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
Topic 20: first local place princess british peace white saturday taken exp
Topic 21: part say sunday members year leaders church days
Topic 22: former ago n't three century rome funeral group year saying led f
Topic 23: france home news south whose president work east take first both
Topic 24: four week throne saying became former died members camilla made l
Topic 25: statement french week thursday war christian born vatican house h
Topic 26: minister prime expected group officials union died children times
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 tu
Topic 30: day friday party private wednesday later british officials former
Topic 31: against
Topic 32: city international good prize second won paris take since died ho
Topic 33: times children world
Topic 34: paris work
Topic 35: michael king paul local father political show ceremony next part
Topic 36: rights government prize service million become officials held pol
Topic 37: television reporters show n't political times clinton several off
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.

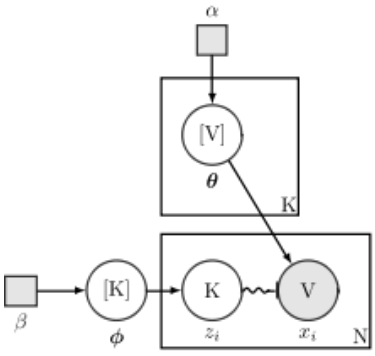

Figure 9. Finite K-sized mixture model currently implemented. θ is the parameter for every cluster component, represented from a categorical draw of over all genes. z_i is the cluster assignment, and ϕ is the distribution of clusters. As usual, α, β are hyper-parameters.