Single-cell network mixed-membership community detection by XXX

Sishir Subedi* Yongjin P. Park†

05:37:45 PM, Mar 02, 2022

Background

The advancement in single-cell RNA-sequencing (scRNA-seq) has emerged as a new frontier in transcriptomics and contributed to our understanding of complex disease biology. The ability to quantify gene expression levels at a single-cell resolution provides a framework to uncover novel cell types, interactions, and dynamics of cellular systems during disease progression.[Nomura, 2021] There are numerous popular tools to analyze gene expression data from single-cell. Most of these tools incorporate a common workflow that includes data normalization, filtering, and representation in lower dimensions for various downstream analyses such as clustering, differential expression, and cell type identification.[Zappia and Theis, 2021] This multi-step process has limitations (TODO:explain), and efforts are ongoing to develop a streamlined and robust computational method to model gene expression levels directly from raw count data.

Autoencoder is an unsupervised machine learning method based on neural networks architecture and has been utilized in many areas of single-cell analysis, such as denoising and clustering.[Eraslan et al., 2019, Geddes et al. [2019]]. These studies have shown that autoencoder-based techniques capture the essential biological signals from sparse and heterogeneous single-cell data by efficiently representing it in lower dimensions. TODO: recent studies related to raw count data, other methods and use of vae

*Bioinformatics Program, The University of British Columbia

†Department of Statistics, The University of British Columbia, ypp@stat.ubc.ca

TODO: another paragraph ...In this study...expand lda, include vae, identify lr interactions

Results

Discussion

Conclusions

Methods

Datasets

- toy data?
- breast cancer - Chung et al. [2017]
- peripheral blood mononuclear cells (pbmc) - Freytag et al. [2018]
- tcells - Zheng et al. [2021]

The Embedded Topic Model

The ETM model extends on the ideas of LDA. Consider a sample of cells $c_1, ..., c_N$ and a list of genes $g_1, ..., g_D$, where $x_{c1}, x_{c2}, ..., x_{cD}$ are raw count data for D genes in cell c. The model represents each cell in terms of K latent topics and each topic is a full distribution over the genes. In LDA, topic proportion θ_c for cell c and topic distribution over genes β_k for topic k are drawn from Dirichlet distribution with fixed model hyperparameters. In ETM, for the topic proportion Dirichlet distribution is replaced with logistic normal distribution, and β_k topic distribution over genes uses softmax function with model hyperparameters α_k.

$$\begin{aligned} \delta_c \sim LN(0, I); \theta_c = softmax(\delta_c) \\ \beta_k = softmax(\alpha_k) \end{aligned} \tag{1}$$

The marginal likelihood

The parameters of ETM model are the topic embeddings $\beta_{1:K}$. The marginal likelihood of cells is given as,

$$
\begin{aligned}
L(\beta) &= \sum_n log\ p(c_n \mid \beta) \\
p(c_n \mid \beta) &= \int p(\delta_c) \prod_d p(x_{cd} \mid \delta_c, \beta) d\delta_c \\
p(x_{cd} \mid \delta_c, \beta) &= \sum_k \theta_{ck} \beta_{k,x_{cd}}
\end{aligned}
\tag{2}
$$

Here, θ_{ck} is topic proportion transformed using softmax fuction over δ_c for cell c, and β_k is distribution over genes induced by topic embeddings α_k. The marginal likelihood of each cell is an intractable problem because it involves integral over the topic proportion. Variational inference techniques can be used to approximate this type of intractable integrals.

Let $p_\theta(z \mid x)$ be a true posterior and $q_\phi(z \mid x)$ be an approximate posterior.

$$
\begin{aligned}
D_{KL}(q_\phi \mid\mid p_\theta) &= E_{q_\phi}[log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z \mid x)] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ \frac{p_\theta(z, x)}{p_\theta(x)}] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z, x)] + E_{q_\phi}[log\ p_\theta(x)] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z, x)] + \int q_\phi(z \mid x) log\ p_\theta(x) dz \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z, x)] + log\ p_\theta(x) \int q_\phi(z \mid x) dz \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z, x)] + log\ p_\theta(x) \\
log\ p_\theta(x) &= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z, x)] + D_{KL}(q_\phi \mid\mid p_\theta)
\end{aligned}
\tag{3}
$$

Here, $D_{KL}(q_\phi \mid\mid p_\theta)$ is intractable but it is always ≥ 0, we can use this property to remove D_{KL} from the equation and the marginal log likelihood $log\ p_\theta(x)$ will be at least $\geq -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z, x)]$. Since this term is the lower bound on the evidence,it is called as Evidence Lower Bound(ELBO). We maximize the marginal log likelihood by maximizing the ELBO and indirectly minimize the KL divergence.

$$
\begin{aligned}
ELBO &= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z, x)] \\
&= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(x \mid z)] + E_{q_\phi}[log\ p_\theta(z)] \\
&= E_{q_\phi}[log\ p_\theta(x \mid z)] - E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z)] \\
ELBO &= E_{q_\phi}[log\ p_\theta(x \mid z)] - E_{q_\phi}[log \frac{q_\phi(z \mid x)}{p_\theta(z)}]
\end{aligned} \tag{4}
$$

Here, $E_{q_\phi}[log\ p_\theta(x \mid z)]$ is an expected reconstruction error and $E_{q_\phi}[log \frac{q_\phi(z|x)}{p_\theta(z)}]$ is KL Divergence between approximate posterior and the prior.

Reconstruction error

Let x_{cd} be count data for d^{th} gene in cell c and P_{cd} be probability of observing x_{cd} count data. Then the likelihood of observing count data for all D genes in cell c is $\prod_d^D P_{cd}^{x_{cd}}$ and log-likelihood is $\sum_d^D x_{cd} log(P_{cd})$.

Approximate posterior and prior

The approximate posterior $q_\phi(z \mid x)$ is a Gaussian variational distribution $q(\delta_c; c_n, v) = N(\mu, \Sigma)$ whose mean and variance are constructed form a neural network parameterized by v. The network takes raw count data of a cell c_n for D genes and outputs a mean and variance of δ_c. The prior $p_\theta(z) = N(0, I)$. The KL divergence between these two form of Gaussians exist in closed form and given as-

$$
\begin{aligned}
D_{KL}(q_\phi \mid\mid p_\theta) &= E_{q_\phi}[log \frac{q_\phi(z \mid x)}{p_\theta(z)}] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z)] \\
&= ....TODO \\
&= 1/2 \sum_d (1 + log(\Sigma) - \mu^2 - \Sigma)
\end{aligned} \tag{5}
$$

How to construct a feature-incidence matrix

- Train ETM model

Algorithm 1 ETM Algorithm

```
Initialize model parameters $\alpha$,$v$
for i in 1,2,... do
    Compute $\beta_k$ = softmax($\alpha_k$) for each topic k
    Choose a minibatch $C$ of cells
    for c in $C$ do
        Get raw count of $D$ genes from cell c as $x_c$
        Compute $\mu_c = NN(x_c; v_\mu)$
        Compute $\Sigma_c = NN(x_c; v_\Sigma)$
        Sample $\delta_c \sim LN(\mu_c, \Sigma_c)$
        Compute $\theta_c$ = softmax($\delta_c$)
        for d genes in cell c do
            Compute $p(x_{cd} \mid \theta_c, \beta) = \theta_c^T \beta_{x_{cd}}$
        end for
    end for
    Calculate the ELBO, gradient, and update parameters
end for
```

- Use latent dimension to find neighbouring cells
- For each neighbourhood, construct an interaction matrix where rows are neighbouring cell pairs and edges are ligand-receptor pairs from known database
- calculate ligand-receptor interaction score

$$f(c_i, c_j, l_x, r_y) = max(e_{c_i l_x} \times e_{c_j r_y}, e_{c_i r_y} \times e_{c_j l_x})$$

$e_{c_i l_x}$ is expression of ligand x in cell i, $e_{c_j r_y}$ is expression of receptor y in cell j, c_i and c_j are neighbouring cells from ETM model, and l_x and r_y are ligand-receptor pairs from known database

- X_{gi}

Clustering the rows of an incidence matrix

Notations

- Y_{eg}: a feature g's contribution to an edge e, $Y \geq 0$
- Z_{ek}: a latent variable for an edge e
- $p(Z_{ek} = \pi)$, where $\pi = 1$ if and only if the edge e belongs to the cluster k; otherwise, $\pi = 0$

- λ_k : parameter vector for a cluster k

Likelihood

$$\begin{aligned} p(Y, Z|\lambda, \pi) &= \prod_g \sum_k p(y_g, z_g = k) \\ &= \prod_g \sum_k p(y_g \mid z_g = k) p(z_g = k) \\ &= \prod_g \sum_k Poisson(y_g \mid \lambda_k^y) \pi_k \\ &= \prod_g \sum_k \prod_e Poisson(y_{eg} \mid \lambda_{ek}^y) \pi_k \end{aligned} \tag{6}$$

Log-likelihood

$$\begin{aligned} log(p(Y, Z|\lambda, \pi)) &= \sum_g log(\sum_k p(y_g, z_g = k)) \\ &\geq \sum_g \sum_k q(z_g = k) log(\frac{p(y_g, z_g = k)}{q(z_g = k)}) \\ &= \sum_g \sum_k q(z_g = k) log(p(y_g, z_g = k)) - q(z_g = k) log(q(z_g = k)) \end{aligned} \tag{7}$$

EM algorithm (like forward-backward of HMM):

1. Step 1. Estimate z given λ

2. Step 2. Estimate λ given z

References

Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications*, 8(1):1–12, 2017.

Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14, 2019.

Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7, 2018.

Thomas A Geddes, Taiyun Kim, Lihao Nan, James G Burchfield, Jean YH Yang, Dacheng Tao, and Pengyi Yang. Autoencoder-based cluster ensembles for single-cell rna-seq data analysis. *BMC bioinformatics*, 20(19):1–11, 2019.

Seitaro Nomura. Single-cell genomics to understand disease pathogenesis. *Journal of Human Genetics*, 66(1):75–84, 2021.

Luke Zappia and Fabian J Theis. Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. *Genome biology*, 22(1):1–18, 2021.

Liangtao Zheng, Shishang Qin, Wen Si, Anqiang Wang, Baocai Xing, Ranran Gao, Xianwen Ren, Li Wang, Xiaojiang Wu, Ji Zhang, et al. Pan-cancer single-cell landscape of tumor-infiltrating t cells. *Science*, 374(6574):abe6474, 2021.