Clustering Single Cells with Noisy Observations

February 4, 2018

1 Introduction

2 Method

2.1 Approach 1

Based on *Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning* [1].

Our goal is to minimize the following in terms of S, L, and w. S is the similarity matrix that we hope to be block diagonal, K is a kernel function differently parametrized for each l, w is the weight of each kernel, and Q is a set of features which we have prior knowledge that they are important in differentiating the samples into clusters of our interest.

$$\min_{S,L,w} -\sum_{i,j,l} w_l K_l(c_i, c_j) S_{i,j} + \beta \|S\|_F^2 + \gamma tr(L^T(I_N - S)L) + \rho \sum_l w_l log w_l - \sum_{q \in \mathcal{Q}} \frac{f_q^T (I - S) f_q}{f_q^T f_q} \tag{1}$$

$$\text{such that } \sum_l w_l = 1,\ w_l \geq 0,\ \sum_j S_{ij} =,\ S_{ij} \geq 0,\ L^T L = I_C$$

As mentioned in [1], although the optimization problem formulated above is nonconvex, the objective function for each variable conditional on the other two variables being fixed is convex.

The original problem from [1] is the same except the last term about the prior knowledge Q. Only updating the S is changed.

- **update S given L and w**

$$\max_S \sum_{i,j,l} w_l K_l(c_i, c_j) S_{ij} + -\beta \|S\|_F^2 - \gamma tr(L^T I L - L^T S L) - \sum_{q \in \mathcal{Q}} \frac{f_q^T S f_q}{f_q^T f_q} \tag{2}$$

$$= \max_S \sum_{i,j} \left(\sum_l (w_l K_l(c_i, c_j)) - \gamma (L L^T)_{ij} - \sum_{q in \mathcal{Q}} \frac{f_{qi} f_{qj}}{\|f_q\|^2} \right) S_{ij} - \beta \|S\|_F^2 \tag{3}$$

$$\text{subject to } \sum_j S_{ij} = 1 \text{ and } S_{ij} \geq 0 \text{ for all } (i,j)$$

The first summation term in the objective as well as constraints are all linear, and the second summation in the objective is a simple quadratic form that can be solved in polynomial time.

- **update L given S and w**

$$max_L tr(L^T(I_N - S)L) \text{ subject to } L^T L = I_C \tag{4}$$

  Then L is the C largest eigenvectors of $I_N - S$.

- **update w given S and L**

$$max_w \sum_l \sum_{i,j} K_l(c_i, c_j)S_{ij} - \rho \sum_l w_l log w_l \tag{5}$$

$$\text{subject to } \sum_l w_l = 1, w_l \geq 0 \tag{6}$$

  This has convex objective and linear constraints and can be solved by any standard convex optimization method.

The last step of similarity enhancement by diffusion and the convergence criterion can both be implemented directly according to the original paper by Wang [1].

2.2 Approach 2

Based on *Robust Multi-View Spectral Clustering via Low-Rank and Sparse Decomposition* [2].

Here, we also will observe and combine different similarity matrices, but unlike the above case where different kernels were used, this case will use different sets of features to compute similarity matrices. The idea is that the averaging of information will naturally help the denoising.

This paper is heavily based on the idea of spectral clustering. From a similarity matrix $S^{(i)}$ for a feature set i, we will construct graph $G^{(i)}$ and corresponding transition matrix $P^{(i)}$. This matrix P will be the main object for minimization, assuming that there exists a true probability transition matrix $\hat{P}$ and the matrices we observe are

$$P^{(i)} = \hat{P} + E^{(i)}$$

where E is the error matrix for each case i. The main optimization goal is

$$\min rank(\hat{P}) + \lambda \sum_i \|E^{(i)}\|_0$$

but since the nonconvexity, we modify the above to

$$\min \|\hat{P}\|_* + \lambda \sum_{i=1}^m \|E^{(i)}\|_1$$

where $*$ is the trace norm. Using Augmented Lagrangian Multiplier, the goal becomes

$$\min_{\hat{P}, Q, E^{(i)}} \|Q\|_* + \lambda \sum_{i=1}^m \|E^{(i)}\|_1$$

such that $i = 1, 2, .., m$, $P^{(i)} = \hat{P} + E^{(i)}$, $\hat{P} \geq 0$, $\hat{P}\mathbf{1} = \mathbf{1}$, $\hat{P} = Q$. The corresponding augmented Lagrangian function is

$$\mathcal{L}(\hat{P}, Q, E^{(i)}) = \|Q\|_* + \lambda \sum_{i=1}^m \|E^{(i)}\|_1 + \sum_{i=1}^m \langle Y^{(i)}, \hat{P} + E^{(i)} - P^{(i)} \rangle \tag{7}$$

$$+ \frac{\mu}{2} \sum_{i=1}^m \|\hat{P} + E^{(i)} - P^{(i)}\|_F^2 + \langle Z, \hat{P} - Q \rangle + \frac{\mu}{2} \|\hat{P} - Q\|_F^2 \tag{8}$$

such that $\hat{P} \geq 0$, $\hat{P}\mathbf{1} = \mathbf{1}$.

Although the notation is different, the transition probability matrix is essentially equal to the weight matrix regarding Laplacian Matrix $L = D - W$, but normalized. Consider this :

$$P_{i,j} = W_{i,j}/Di, i$$

and naturally

$$I_{i,j} = D_{i,j}/D_{i,i}$$

because $D_{i,j}$ is 0 for all $i \neq j$. Therefore, our new Laplacian matrix can be written like $L = I_n - P$

One of the properties of Laplacian is that it has rank $n - k$ where k is the number of connected components. Then, the rank of P is k. In the approach 2, we minimize the trace of P, and therefore we minimize the number of connected component k. In the approach 1 however, we minimize the trace of $L^T(I - S)L$ where L here is not Laplacian but an orthogonal matrix with rank C such that $L^T L = I_C$. Does this make sense? Verify.

Therefore, we can add the extra optimization term

$$\sum_{q \in \mathcal{Q}} \frac{f_q^T (I - \hat{P}) f_q}{\|f_q\|_2^2}$$

We would like to build an optimization problem that accounts for any prior information about important features. This should be implemented before

Questions :

- Multiplier at the penalty for the prior knowledge?
- intuition behind trace norm and the multi-kernel problem. Should I try both?
- Double check the interpretation is correct (that the transition probability matrix is essentially S).

$$\min_X \|X\|_* + \lambda \sum_{i=1}^{m} \|X - \hat{X}^{(i)}\|_1 \text{ such that } X \geq 0, X1 = 1$$

3 KNN-kernel density-based clusteringn for high-dimensional multivariate data

kernel density estimation

Consider $N \times d$ dimensional data. The d-dimensional space can be partitioned into a number of equal bins. Consider the density below:

$$\hat{f}(x) = \frac{1}{NV} \sum_{i=1}^{N} K((x - x_i)./H$$

The size of the bin is given by a scale vector $H = [h_1, .., h_d]$. V is data volume $\prod_{i=1}^{d} h_i$.

References

[1] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods*, 14(4):414–416, 2017.

[2] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, pages 2149–2155, 2014.