Clustering Single Cells with Noisy Observations

February 19, 2018

Problem Set up

Based on Robust Multi-View Spectral Clustering via Low-Rank and Sparse Decomposition [1].

We observe a noisy feature matrix with n rows and p columns where n is the number of cells and p is the number of predictors (genotype, gene expression level, etc). We aim to recover the true similarity matrix of the cells that is **low rank**. We divide the predictors into m groups and for each group we compute the similarity matrix $S^{(i)}$ for $i = 1, \cdots, m$. Adopting the idea of spectral clustering, for each similarity matrix $S^{(i)}$ for the i'th feature group, we construct a graph $G^{(i)}$ and the corresponding transition matrix $P^{(i)}$.

We assume that these transition probability matrices are the superpositions of the true low rank probability transition matrix $\hat{P}$ and the sparse error matrix $E^{(i)}$.

In addition, we add a separate term on the optimization objective where users can add prior information. For example, there exist lists of genes that play a significant role in classifying the cells. Calling this set $\mathcal{Q}$, for each $q \in \mathcal{Q}$, we would like to maximize the Laplacian

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

From below, we assume that all the feature vectors have been normalized to sum of squares 1, and only use

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

Therefore, under this setting, our optimization goal is

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

but we actually solve below with convex relaxation.

$$\min_{\hat{P}, E^{(i)}} \|\hat{P}\|_* + \lambda \sum_{i=1}^{m} \|E^{(i)}\|_1 \text{ such that } P^{(i)} = \hat{P} + E^{(i)}, \hat{P} \geq 0, \hat{P}\mathbf{1} = \mathbf{1}, \hat{\mathbf{P}} = \mathbf{Q} \text{ for } \mathbf{i} = \mathbf{1}, \cdots, \mathbf{m}$$

We introduce an auxiliary variable Q for for ADMM method. The corresponding augmented Lagrangian function is

$$\begin{aligned}\mathcal{L}(\hat{P}, Q, E^{(i)}) = \|Q\|_* + \lambda \sum_{i=1}^{m} \|E^{(i)}\|_1 + \sum_{i=1}^{m} \langle Y^{(i)}, \hat{P} + E^{(i)} - P^{(i)} \rangle \\ + \frac{\mu}{2} \sum_{i=1}^{m} \|\hat{P} + E^{(i)} - P^{(i)}\|_F^2 + \langle Z, \hat{P} - Q \rangle + \frac{\mu}{2} \|\hat{P} - Q\|_F^2\end{aligned}$$

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

We add the information of prior information in the objective like the following.

$$\min_{\hat{P}, E^{(i)}} \|\hat{P}\|_* + \lambda \sum_{i=1}^{m} \|E^{(i)}\|_1 + \sum_{q \in \mathcal{Q}} \rho_q f_q^T (I - P) f_q$$

such that $i = 1, 2, .., m,\ P^{(i)} = \hat{P} + E^{(i)},\ \hat{P} \geq 0,\ \hat{P}\mathbf{1} = \mathbf{1},\ \hat{P} = Q$.
Following the paper's algorithm directly, we set up the conditional updates of Q, $E^{(i)}$, and $\hat{P}$

Update Q

$$Q = argmin_Q \|Q\|_* + \frac{\mu}{2}\|Q^T - \hat{P}^T - \frac{Z^T}{\mu}\|_F^2$$

Solve this with Singular Value Threshold method.

Update $E^{(i)}$

$$E^{(i)} = argmin_{E^{(i)}} \lambda \|E^{(i)}\|_1 + \frac{\mu}{2}\|E^{(i)} - (P^{(i)} - \hat{P} - \frac{Y^{(i)}}{\mu})\|_F^2$$

This has closed form solution with soft thresholding function.

Update $\hat{P}$

$$\begin{aligned}\hat{P} = argmin_{\hat{P}} \frac{\mu}{2} \sum_{i=1}^{m} \|\hat{P} + E^{(i)} - P^{(i)} + \frac{Y^{(i)}}{\mu}\|_F^2 \\ + \frac{\mu}{2}\|\hat{P} - Q + \frac{Z}{\mu} + \sum_{q \in \mathcal{Q}} f_q f_q^T / \mu\|_F^2\end{aligned}$$

For notational convenience, introduce C:

$$C = \frac{1}{m+1}\left(Q - \frac{Z}{\mu} - \sum_{q \in \mathcal{Q}} \rho_q \frac{f_q f_q^T}{\mu} - \sum_{i=1}^{m}\left((P^{(i)} - E^{(i)} - \frac{Y^{(i)}}{\mu}\right)\right)$$

and we solve the following optimization problem

$$\hat{P} = argmin_{\hat{P}} \frac{1}{2}\|\hat{P} - C\|_F^2 \text{ such that } \hat{P} \geq 0, \hat{P}\mathbf{1} = \mathbf{1}$$

which has convex function with linear constraints. This can be solved using projection algorithm.

Passing Evaluation

The rand index is good - not exceptional, just as expected - and NMI is significantly worse than the recent paper. There are many ways for improvement: (1) adaptively optimize the hyperparameters, (2) adopt network diffusion, and (3) enforce the rank of the similarity matrix as the user-defined number of clusters.

(2). I read the original paper, but I couldn't really see how this can be applied to our problem (nevertheless interesting). I read the supplementary materials, and I read the codes. I think the gist is that using k-nearest-neighbors reduces the noise in the similarity matrix by making it sparse. However, our original goal does not include the sparsity of the matrix $\hat{P}$. We only need the low-rank condition. Implementing network diffusion in $\hat{P}$ led to inflated $E^{(i)}$, and the clustering result was inferior to the original algorithm without the network diffusion.

(3) Now I plan to implement the rank-enforcing structure in our optimization objective. The Nature paper seems to have a typo - the term

$$tr(L^T(I-S)L)$$

does not *minimize* the rank of S. Rather, it puts a lower bound for the rank of $S: rank(S) \geq C$. Since they minimize $\|S\|_F^2$, they are pushing *up* the rank of S. In our context, however, we'd like to have a rank constraint $rank(\hat{P}) \leq C$. That means, we need the term

$$\min_{Q,L} \|L^T Q L\|_*$$

such that $L_T L = I_{n-C}$

Rank enforcing

$$\begin{aligned}\mathcal{L}(\hat{P}, Q, E^{(i)}, L) = \|L^T Q L\|_* + \lambda \sum_{i=1}^{m} \|E^{(i)}\|_1 + \sum_{i=1}^{m} \langle Y^{(i)}, \hat{P} + E^{(i)} - P^{(i)} \rangle \\ + \frac{\mu}{2} \sum_{i=1}^{m} \|\hat{P} + E^{(i)} - P^{(i)}\|_F^2 + \langle Z, \hat{P} - Q \rangle + \frac{\mu}{2} \|\hat{P} - Q\|_F^2\end{aligned}$$

such that $\hat{P} \geq 0$, $\hat{P}\mathbf{1} = \mathbf{1}$, and $L^T L = I_{n-C}$ where C is the number of clusters.

Updating $\hat{P}$ and $E^{(i)}$ is the same as before. When updating L, we simply take svd of $Q - I$ and take C first eigenvectors corresponding to the largest eigenvalues. When updating Q, we solve below.

Updating Q :

$$\begin{aligned}Q &= \underset{Q}{\operatorname{argmin}} \|L^T Q L\|_* + \langle Z, \hat{P} - Q \rangle + \frac{\mu}{2} \|\hat{P} - Q\|_F^2 \\ &= \underset{Q}{\operatorname{argmin}} \|L^T Q L\|_* + \frac{\mu}{2} tr\left(Q^T Q - \left(P^T + \frac{Z^T}{\mu}\right) Q\right) \\ &= \underset{Q}{\operatorname{argmin}} \|L^T Q L\|_* + \frac{\mu}{2} \left\| Q - P - \frac{Z}{\mu} \right\|_F^2\end{aligned}$$

Adaptively determining hyper-parameters

Ways to decide the penalty λ, starting μ, the number of splits m, and σ for the kernel.

Notes

- The Laplacian matrix has rank $n-k$ where k is the number of connected components in the graph.
- The smallest eigenvector of Laplacian matrix is always 0, and its corresponding eigenvector is always $1/\sqrt{n}$. Its second smallest eigenvalue's corresponding eigenvector is the one that clusters the graph components.

- Batzoglou paper's optimization objective clearly has a constraint $\sum_j S_{ij} = 1$ but the algorithm's output doesn't satisfy this. Probably due to the network diffusion - knn procedure. It didn't normalize S after the network diffusion step.

- There's a typo in Batzoglou's paper. The rank minimization term does not minimize the rank of S, it minimizes the rank of $I - S$. Think about the case when S has rank 1 - L will be picked to be a vector e_1, and when it's again turn to pick S with fixed L, it's okay as long as S has 1 in its first singular value, and have whatever (zero or nonzero) singular values for the rest. In other words, it's constraining $rank(S) \geq C$, not $rank(S) \leq C$.

- general spectral clustering : the 2017 paper forces S to be symmetric at every iteration, and violates their initial constraint $S\mathbf{1} = \mathbf{1}$. Our algorithm makes $\hat{P}$ symmetric only after the convergence. Should this matter?

- sparsity constraint? would it help?

New model with weighted least squares

Define S as our **true** similarity matrix. We assume S is both low rank and sparse as well as positive semi-definite. We observe multiple noisy observations $P^{(i)}$ where P is computed via multiple kernels using the single cell sequencing data. Then, we model the data like the following.

$$P^{(i)} = S + E^{(i)}, \quad E^{(i)}_{jk} \sim N(0, \sigma_i^2)$$

That means, elements of the error matrix for kernel i follows a normal distribution with different variance σ_i. Then we can build our optimization problem like the following :

$$\hat{S} = \underset{S}{\operatorname{argmin}}\, \ell(S, P) + \tau \|S\|_* + \gamma \|S\|_1$$

where $\ell(S, P) = \sum_{i=1}^m \frac{1}{\sigma_i^2} \|S - P^{(i)}\|_F^2$ is the differentiable convex loss function based on the observed probability transition matrix P, and we have further simplex constraint on S : $S\mathbf{1} = \mathbf{1}$, $S \geq 0$. Lastly, we have a constraint on the σ so that the sum of the loss has weight 1: $\sum_{i=1}^m \frac{1}{\sigma_i^2} = 1$. Re-naming $\frac{1}{\sigma_i^2}$ with λ_i,

$$\hat{S} = argmin_S \sum_{i=1}^m \lambda_i^2 \|S - P^{(i)}\|_F^2 + \tau \|S\|^* + \gamma \|S\|_1, \text{ where } S\mathbf{1} = \mathbf{1},\ S \geq 0, \sum_{i=1}^m \lambda_i = 1$$

Using the incremental proximal descent, we can use the following algorithm.

Algorithm 1 Incremental Proximal descent

Init ialize $S = \frac{1}{m} \sum_{i=1}^m P^{(i)}$
Repeat until convergence
update S : $\operatorname{argmin} \ell(S, P)$ s.t. $S1 = 1, S \geq 0$ (Duchi 2008)
update $S = \operatorname{prox}_{\theta\tau\|\cdot\|_*}(S)$
update $S = \operatorname{prox}_{\theta\gamma\|\cdot\|_1}(S)$
update λ_i : $\frac{\|E^{(i)}\|_F^2}{\sum_j \|E^{(j)}\|_F^2}$

Or, using augmented Lagrangian, we can solve the following.

$$
\begin{aligned}
\mathcal{L} = \tau\|Q\|_* + \gamma\|R\|_1 + \sum_{i=1}^{m} \lambda_i^2 \|E^{(i)}\|_F^2 & \\
+ \langle Y, S - Q\rangle + \frac{\mu}{2}\|S - Q\|_F^2 + \langle Z, S - R\rangle + \frac{\mu}{2}\|S - R\|_F^2 & \\
+ \sum_{i=1}^{m} \langle W^{(i)}, S + E^{(i)} - P^{(i)}\rangle + \frac{\mu}{2}\|S + E^{(i)} - P^{(i)}\|_F^2 & \\
\text{s.t. } S1 = 1, S \geq 0, \sum_i \lambda_i = 1 &
\end{aligned}
$$

In order to solve this, we use the following updates for ADMM algorithm.

- update Q

$$\underset{Q}{\operatorname{argmin}}\, \tau\|Q\|_* + \frac{\mu}{2}\|Q - S - \frac{Y}{\mu}\|_F^2$$

Take SVD of $\mathcal{S} + \frac{Y}{\mu}$, and the solution for Q is

$$Q = U\mathcal{S}_{\frac{\tau}{\mu}}(\Sigma)V^T$$

where $\mathcal{S}_{\frac{\tau}{\mu}}(\Sigma)$ is soft-thresholding the singular values at $\frac{\tau}{\mu}$.

- update R

$$\underset{R}{\operatorname{argmin}}\, \gamma\|R\|_1 + \frac{\mu}{2}\|R - S - \frac{Z}{\mu}\|_F^2$$

The solution for R is

$$R = \mathcal{S}_{\frac{\gamma}{\mu}}(S + \frac{Z}{\mu})$$

- update S

$$S = \underset{S}{\operatorname{argmin}}\, \frac{(m+2)\mu}{2}\|S - \frac{1}{m+2}(R + Q + \sum_i (E^{(i)} - P^{(i)} - \frac{W^{(i)}}{\mu}) - \frac{Y}{\mu} - \frac{Z}{\mu}\|_F^2 \text{ such that } S1 = 1, S \geq 0$$

$$\underset{S}{\operatorname{argmin}}\, \frac{1}{2}\|S - C\|_F^2 \text{ s.t. } S \geq 0, S1 = 1 \text{ where } C = \frac{1}{m+2}\left(R + Q + \sum_i (E^{(i)} - P^{(i)} - \frac{W^{(i)}}{\mu}) - \frac{Y}{\mu} - \frac{Z}{\mu}\right)$$

We can then use Duchi 2008 algorithm for proximal operator with simplex constraint.

- update $E^{(i)}$

$$
\begin{aligned}
E^{(i)} &= \underset{E^{(i)}}{\operatorname{argmin}}\, \lambda_i^2\|E^{(i)}\|_F^2 + tr(W^{(i)T}E^{(i)}) + \frac{\mu}{2}tr(E^{(i)T}E^{(i)} + E^{(i)T}(S - P^{(i)})) \\
&= \underset{E^{(i)}}{\operatorname{argmin}}\, tr\left((\lambda_i^2 + \frac{\mu}{2})E^{(i)T}E^{(i)} + \frac{\mu}{2} \cdot 2E^{(i)T}\left(\frac{W^{(i)}}{\mu} + S - P^{(i)}\right)\right) \\
&= \underset{E^{(i)}}{\operatorname{argmin}}\left(\lambda_i^2 + \frac{\mu}{2}\right)\left\|E^{(i)} - \frac{\frac{\mu}{2}}{\lambda_i^2 + \frac{\mu}{2}}\left(P^{(i)} - S - \frac{W^{(i)}}{\mu}\right)\right\|_F^2
\end{aligned}
$$

Therefore, $E^{(i)} = \frac{1}{\lambda_i^2 + \frac{\mu}{2}}\left(P^{(i)} - S - \frac{W^{(i)}}{\mu}\right)$

- update λ_i

$$\lambda_i = \frac{\frac{1}{\|E^{(i)}\|_F^2}}{\sum_j \frac{1}{\|E^{(j)}\|_F^2}}$$

Newer model that adaptively decides λ

Reference : Scaled Sparse Linear Algebra (Tingni Sun and Cunhui Zhang)

Consider two penalty functions:

$$L_\lambda(\beta) = \frac{\|Y - X\beta\|_2^2}{2n} + \lambda\|\beta\|_1 \tag{1}$$

$$L_{\lambda_0}(\beta, \sigma) = \frac{\|Y - X\beta\|_2^2}{2n\sigma} + \lambda_0\|\beta_1\| + \frac{(1-a)\sigma}{2} \tag{2}$$

The paper's gist is that the following algorithm leads to the solution of minimizing the Equation (2).

$$\hat{\sigma} = \frac{\|Y - X\hat{\beta}^{old}\|_2}{((1-a)n)^{1/2}} \tag{3}$$

$$\lambda = \hat{\sigma}\lambda_0 \tag{4}$$

$$\hat{\beta} = \beta(\lambda) \text{ from lasso path} \tag{5}$$

The new β computed from the lasso path minimizes (2) and new $\hat{\sigma}$ does as well. (The derivation is simple.)
Applying the above, consider the following optimization function:

$$\underset{S}{\operatorname{argmin}}\, \lambda_0\|S\|_1 + \sum_{i=1}^{m} \frac{\|S - P^{(i)}\|_F^2}{2n\sigma_i} + \frac{(1-a)\sigma_i}{2} \tag{6}$$

This is almost identical to the scaled sparse linear regression except that there are multiple σ's. Alternatively updating σ_1, ..., σ_m, and S will lead to the desired result, though. Now, consider the following with the nuclear norm.

$$\underset{S}{\operatorname{argmin}}\, \tau\|S\|_* + \lambda_0\|S\|_1 + \sum_{i=1}^{m} \frac{\|S - P^{(i)}\|_F^2}{2n\sigma_i} + \frac{(1-a)\sigma_i}{2} \tag{7}$$

For this, a general idea is to create $m+1$ auxiliary variables. One for the trace norm and m for the 1-norm. That means, for each i, we will make auxiliary variable $R^{(i)}$

Consider the following optimization goal :

$$\mathcal{L} = \tau\|Q\|_* + \sum_{i=1}^{m} \frac{\|E^{(i)}\|_F^2}{2n\sigma_i} + \frac{\sigma_i}{2} + \langle Y, S - Q\rangle + \frac{\mu}{2}\|S - Q\|_F^2$$
$$+ \sum_{i=1}^{m}\langle Z^{(i)}, P^{(i)} - S - E^{(i)}\rangle + \sum_{i=1}^{m}\frac{\mu}{2}\|P^{(i)} - S - E^{(i)}\|_F^2$$

- update Q

$$\begin{aligned} Q &= \arg\min_Q \tau\|Q\|_* + tr(-Q^TY - \frac{\mu}{2}2Q^TS + \frac{\mu}{2}Q^TQ) \\ &= \arg\min_Q \tau\|Q\|_* + \frac{\mu}{2}tr(Q^TQ - 2Q^T(S + \frac{Y}{\mu})) \\ &= \arg\min_Q \tau\|Q\|_* + \frac{\mu}{2}\left\|Q - \left(S + \frac{Y}{\mu}\right)\right\|_F^2 \end{aligned}$$

  Use singular value threshold to update Q.

- update S

$$\begin{aligned} S &= \arg\min_S tr(Y^TS) + \frac{\mu}{2}tr(S^TS - 2S^TQ) - \sum_{i=1}^{m} Z^{(i)T}S + \sum_{i=1}^{m}\frac{\mu}{2}tr(S^TS - P^{(i)T}S + E^{(i)T}S) \\ &= \arg\min_S \frac{\mu}{2}tr\left((m+1)S^TS - \sum_{i=1}^{m}(P^{(i)} - E^{(i)})^TS + \left(\frac{Y}{\mu} - \sum_{i=1}^{m}\frac{Z^{(i)}}{\mu} - Q\right)^T S\right) \end{aligned}$$

- update E
- update σ

New model

Consider this :

$$\min_{S,\sigma_i} \sum_{i=1}^{m} \frac{\|S - P^{(i)}\|_F^2}{2n\sigma_i} + \frac{\sigma_i}{2}$$
$$\text{such that } rank(S) \le r$$

Then we can construct the augmented lagrangian like this:

$$\mathcal{L}(Q^{(i)}, \sigma, S) = \sum_{i=1}^{m} \frac{\|Q^{(i)} - P^{(i)}\|_F^2}{2n\sigma_i} + \frac{\sigma_i}{2} + \sum_{i=1}^{m}\langle Y^{(i)}, S - Q^{(i)}\rangle + \frac{\mu}{2}\|S - Q^{(i)}\|_F^2$$

with the rank constraints of

$$rank(Q^{(i)}) \le r, rank(S) \le r$$

The updating scheme is similar.

- update $Q^{(i)}$:

$$
\begin{aligned}
&\underset{Q^{(i)}}{\operatorname{argmin}} \frac{tr(Q^{(i)T}Q^{(i)} - 2P^{(i)T}Q^{(i)})}{2\sigma_i n} - tr(Y^{(i)T}Q^{(i)}) + \frac{\mu}{2} tr(Q^{(i)T}Q^{(i)} - 2S^T Q^{(i)}) \\
=&\underset{Q^{(i)}}{\operatorname{argmin}}\, tr\left(\left(\frac{\mu}{2} + \frac{1}{2\sigma_i n}\right) Q^{(i)T}Q^{(i)} - \left(\mu S + \frac{P^{(i)}}{\sigma_i n} + Y^{(i)}\right)^T Q^{(i)}\right) \\
=&\underset{Q^{(i)}}{\operatorname{argmin}} \left(\frac{\mu\sigma_i n + 1}{2\sigma_i n}\right) \|Q^{(i)} - \frac{\sigma_i n}{\mu\sigma_i n + 1}\left(\mu S + \frac{P^{(i)}}{\sigma_i n} + Y^{(i)}\right)\|_F^2
\end{aligned}
$$

- update S:

$$
\begin{aligned}
&\underset{S}{\operatorname{argmin}} \sum_{i=1}^{m} tr\left(Y^{(i)T}S + \frac{\mu}{2} S^T S - \mu S^T Q^{(i)}\right) \\
=&\underset{S}{\operatorname{argmin}}\, tr\left(\frac{\mu m}{2} S^T S - \sum_i (\mu Q^{(i)} - Y^{(i)})^T S\right) \\
=&\underset{S}{\operatorname{argmin}} \frac{\mu m}{2} \|S - \frac{1}{\mu m} \sum_i (\mu Q^{(i)} - Y^{(i)})\|_F^2
\end{aligned}
$$

Now, add sparsity constraint to the model above.

$$
\min_{S,\sigma_i} \sum_{i=1}^{m} \frac{\|S - P^{(i)}\|_F^2}{2n\sigma_i} + \|S\|_1 + \frac{\sigma_i}{2}
$$

$$
\text{such that } rank(S) \leq r
$$

Then we can construct the augmented lagrangian like this:

$$
\mathcal{L}(Q^{(i)}, \sigma, S) = \sum_{i=1}^{m} \frac{\|Q^{(i)} - P^{(i)}\|_F^2}{2n\sigma_i} + \frac{\sigma_i}{2} + \sum_{i=1}^{m} \langle Y^{(i)}, S - Q^{(i)} \rangle + \frac{\mu}{2} \|S - Q^{(i)}\|_F^2
$$

ADDING w version 1

$$\min_S \sum_i w_i \|P^{(i)} - S\|_F^2 + \sum_i w_i^2, \quad rank(S) \leq r, \sum_i w_i = 1$$

Now using ADMM method,

$$\mathcal{L}(Q, S, w_i) = \sum_i \frac{1}{2} w_i \|P^{(i)} - Q^{(i)}\|_F^2 + \sum_i \langle Y^{(i)}, S - Q^{(i)} \rangle + \frac{\mu}{2} \|S - Q^{(i)}\|_F^2 + \frac{1}{2} \sum_i w_i^2$$

$$\begin{aligned}
\arg\min_{Q^{(i)}} \mathcal{L} &= \arg\min_{Q^{(i)}} \frac{1}{2} tr \left(Q^{(i)T} Q^{(i)} w_i - 2P^{(i)T} Q^{(i)} w_i - 2Y^{(i)T} Q^{(i)} + \mu Q^{(i)T} Q^{(i)} - \mu 2 S^T Q^{(i)} \right) \\
&= \arg\min_{Q^{(i)}} \frac{1}{2} tr \left((w_i + \mu) Q^{(i)T} Q^{(i)} - 2(P^{(i)} w_i + Y^{(i)} + \mu S)^T Q^{(i)} \right) \\
&= \arg\min_{Q^{(i)}} \frac{w_i + \mu}{2} \|Q^{(i)} - \frac{1}{w_i + \mu} (w_i P^{(i)} + Y^{(i)} + \mu S)\|_F^2 \\
&= \frac{1}{w_i + \mu} (w_i P^{(i)} + Y^{(i)} + \mu S) \\
\arg\min_S \mathcal{L} &= \arg\min_S \frac{\mu}{2} \|S - \frac{1}{m} \sum_{i=1}^m (Q^{(i)} + \frac{Y^{(i)}}{\mu})\|_F^2 \text{ such that rank(S)} \leq \text{r}
\end{aligned}$$

For rank constraint, take the SVD of $\frac{1}{m} \sum_{i=1}^m (Q^{(i)} + \frac{Y^{(i)}}{\mu})$, and truncate after the first r singular values.

$$\frac{1}{m} + \left(\frac{1}{m} \sum_j \left(\|P^{(j)} - Q^{(j)}\|_F^2 \right) - \|P^{(i)} - Q^{(i)}\|_F^2 \right)$$

ADDING w version 2

$$\min_S \sum_i \left(w_i \frac{\|S - P^{(i)}\|_F^2}{2n\sigma_i} + w_i^2 + \frac{\sigma_i}{2} \right), \quad rank(S) \le r, \sum_i w_i = 1$$

Now using ADMM method,

$$\mathcal{L} = \sum_i \frac{1}{2} \frac{w_i}{n\sigma_i} \|Q^{(i)} - P^{(i)}\|_F^2 + \sum_i \langle Y^{(i)}, S - Q^{(i)} \rangle + \frac{\mu}{2} \|S - Q^{(i)}\|_F^2 + \frac{1}{2} \sum_i w_i^2 + \frac{\sigma_i}{2}$$

$$\arg\min_{Q^{(i)}} \mathcal{L} = \frac{n\sigma_i}{w_i + n\sigma_i \mu} \left(\frac{w_i}{n\sigma_i} P^{(i)} + Y^{(i)} + \mu S \right)$$

$$\arg\min_S \mathcal{L} = \arg\min_S \frac{\mu}{2} m \|S - \frac{1}{m} \sum_i (Q^{(i)} - \frac{Y^{(i)}}{\mu})\|_F^2 \text{ such that rank(S)} \le \text{r}$$

So for S, take SVD of $\frac{1}{m} \sum_i Q^{(i)} - \frac{Y^{(i)}}{\mu}$

$$\arg\min_{\sigma_i} \mathcal{L} = \frac{w_i \|Q^{(i)} - P^{(i)}\|_F^2}{2n}$$

$$\arg\min_{w_i} \mathcal{L} = \arg\min_{w_i} \frac{1}{2} \frac{w_i}{n\sigma_i} \|Q^{(i)} - P^{(i)}\|_F^2 + \frac{1}{2} \sum_i w_i^2 - \lambda(w_1 + \cdots + w_m - 1) = 0$$

$$= \frac{1}{m} + \left(\frac{1}{m} \sum_j \frac{\|Q^{(i)} - P^{(i)}\|_F^2}{2n\sigma_j} - \frac{\|Q^{(i)} - P^{(i)}\|_F^2}{2n\sigma_i} \right)$$

This derivation for w_i is clearly wrong because we don't have the constraint that $w_i \ge 0$. Maybe we should go back to $w_i log(w_i)$?

References

[1] 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.