Graph-Sparse Decomposition

Johnny Li

October 2018

1 Graph-Sparse Decomposition

We would like to find a low-rank decomposition of an input matrix $X \in \mathbb{R}^{L \times N}$ with N gene measurements for L samples. We'd also like the components of the decomposition to be informed by a prior knowledge graph, $\mathcal{G}$, such that the support of each component forms a tree with low-cost edges. We formalize this as the following optimization problem:

$$\arg\min_{D,Z} \|X - ZD\|_2^2 + \lambda \sum_{i=1}^{K} c(D_i), \tag{1}$$

$$\arg\min_{U,V} \|X - UV^T\|_F^2 + \lambda \sum_{i=1}^{K} c(U_i), \tag{2}$$

$$f(\mathbf{u}) \xrightarrow{\text{PCST}} \mathbf{u}' \tag{3}$$

$$\arg\min_{T=(V',E')} \pi(\overline{V'}) + c(E') \tag{4}$$

where $D = \{D_i\} \in \mathbb{R}^{K \times N}$ is a dictionary of components such that the support of each D_i forms a tree when projected onto $\mathcal{G}$, and $Z \in \mathbb{R}^{L \times K}$ is the corresponding scores matrix that weights each component. The second term adds a sparsity constraint to the optimization in the form $c(D_i)$, which is the minimum sum of edges used to connect the support of D_i in $\mathcal{G}$ (aka minimum spanning tree). Finally, the λ parameter weights the edge costs and effectively modulates the sparsity of D.

2 Method

To solve this minimization problem, we alternate updates to the score matrix, Z, and the dictionary, D.

First, we fix D and optimize Z. Because D is constant, the sparsity term in (5) can be ignored so that

$$Z^* = \arg\min_{Z} \|X - ZD\|_2^2 = (DD^{\mathrm{T}})^{-1}DX, \tag{5}$$

where Z^* is given by the ordinary least squared (OLS) estimator. Next, we update one dictionary component D_u while keeping the remaining components $\{D_i\}_{i \neq u}$ and Z fixed. Breaking the product ZD into K individual outer products, (5) becomes

$$D_u^* = \arg\min_{D_u} \left\| X - \sum_{i=1}^{K} Z_i D_i^{\mathrm{T}} \right\|_2^2 + \lambda \sum_{i=1}^{K} c(D_i), \tag{6}$$

$$= \arg\min_{D_u} \left\| X' - Z_u D_u^{\mathrm{T}} \right\|_2^2 + \lambda c(D_u), \tag{7}$$

where $X' = X - \sum_{i \neq j} Z_i D_i^{\mathrm{T}}$. In (7), we have ignored all sparsity terms for components that are not D_u because they are fixed. Next we note that, because the edge cost function is invariant given the support of D_u, each element of $D_u = \{D_u(v)\}_N$ must be either 0 or $D_u'(v)$, where

$$D_u' = \arg\min_{D_u} \left\| X' - Z_u D_u^{\mathrm{T}} \right\|_2^2. \tag{8}$$

Again, D_u' can be computed directly via OLS. If we denote X_i' to be the vector of values for gene i across all L samples, we can rewrite our objective as

$$D_u^* = \arg\min_{D_u \in \{0, D_u'(\cdot)\}^N} \sum_{i=1}^{N} \|X_i' - Z_u D_u(i)\|^2 + \lambda c(D_u) \tag{9}$$

$$= \arg\min_{D_u \in \{0, D_u'(\cdot)\}^N} \sum_{i=1}^{N} \left(\|X_i' - Z_u D_u(i)\|^2 - \|X_i'\|^2 \right) + \lambda c(D_u) \tag{10}$$

where we have subtracted the constant value $\sum_N \|X_i'\|^2$ to obtain (10). Now, we set the prize of node i to be the amount of error reduced if $D_u(i)$ is $D_u'(i)$ instead of 0:

$$p(i) = \|X_i'\|^2 - \|X_i' - Z_u D_u(i)\|^2. \tag{11}$$

Note that $p(i) = 0$ if $D_u(i) = 0$, namely if i is not in the support (tree) of D_u. Then we have

$$
\begin{aligned}
D_u^* &= \arg\min_{D_u} \; -\sum_{i=1}^{N} p(i) + \lambda c(D_u) && (12)\\
&= \arg\min_{D_u} \; -\sum_{i \in \mathrm{supp}(D_u)} p(i) + \lambda c(D_u) && (13)\\
&= \arg\min_{D_u} \; \sum_{i \notin \mathrm{supp}(D_u)} p(i) + \lambda c(D_u) && (14)\\
& && (15)
\end{aligned}
$$

So the solution to (6) can be found using a PCST solver.