Mathematics for Data Sciences **April 5, 2016**

Lecture 04. Some remarks on MDS

Instructor: Yuan Yao, Peking University *Scribe: Zhu, Weizhi*

1 Comparison of two kinds of MDS

Given a distance matrix $D = [d_{ij}^2]_{n\times n}$, with $d_{ij} \geq 0, d_{ii} = 0$, and $d_{ij} = d_{ji}$, define $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathbf{T}}$and ,$B = -\frac{1}{2}HDH$.

In this section, we consider two minimization problems of MDS:

$$\min_{Y\in\mathbb{R}^{k\times n}} L_0(Y) := \|Y^TY - B\|_F^2 \tag{1}$$

and

$$\min_{Y_i\in\mathbb{R}^k} L_1(Y) := \sum_{i,j}(\|Y_i - Y_j\|^2 - d_{ij}^2)^2 \tag{2}$$

We have already shown the optimal solution of (1), or the best rank-k approximation,can be obtained by the SVD decomposition of B. Now we focus our attention on how to connect (2) with (1).In fact, from $B = -\frac{1}{2}HDH$, one can shows that $D = \Phi(B)J + J\Phi(B) - 2B$, where $\Phi([a_{ij}]) = diag(a_{ii})$, and $J = [1]$, the matrix of all ones.Thus,

$$\begin{aligned}
\mathcal{L}_1(Y) &= \|(\Phi(Y^TY)J + J\Phi(Y^TY) - 2Y^TY) - (\Phi(B)J + J\Phi(B) - 2B)\|_F^2 \\
&= \|(\Phi(B - Y^TY)J + J\Phi(B - Y^TY)) - 2(B - Y^TY)\|_F^2 \\
&= \|(\Phi(R)J + J\Phi(R)) - 2(R)\|_F^2 \\
&= \|\Phi(R)J + J\Phi(R)\|_F^2 - 4tr(RJ\Phi(R) + \Phi(R)JR) + 4\|R\|_F^2 \\
&= \|\Phi(R)J + J\Phi(R)\|_F^2 + 4\|R\|_F^2
\end{aligned}$$

Where $R =: B - Y^TY$,the $4th$ equality is due to $\|A\|_F^2 = tr(A^TA)$, and the $6th$ equality is due to $JR = RJ = \mathbf{0}$,note that we assume here Y^TY has been centered,that is,it has zero column and row sums.

The first term above,could be further simplified by

$$\begin{aligned}
\|\Phi(R)J + J\Phi(R)\|_F^2 &= n\sum_{i=1}^n r_{ii}^2 + 2(\sum_{i=1}^n r_{ii})(\sum_{j=1}^n r_{jj}) + n\sum_{j=1}^n r_{jj}^2 \\
&= 2n\sum_{i=1}^n r_{ii}^2 + 2(tr(R))^2
\end{aligned}$$

As a result, we have

$$L_1(Y) = 2n\sum_{i=1}^n r_{ii}^2 + 2(tr(R))^2 + 4\|R\|_F^2 \tag{3}$$

Now,we present a specific example to see the difference between the two problems: Suppose Y has SVD decomposition $Y = U\Lambda V^T$, where$\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$. We're looking for optimal in $\tilde{Y} = \tilde{\Lambda}^{\frac{1}{2}} V^T$, where $\Lambda = diag(\tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_n})$, $\tilde{\lambda_i} \geq 0$. (However,readers may not be worried about whether $\tilde{Y}^T\tilde{Y}$ has zero row and column sums. Note that both Y and $\tilde{Y}$ share the property : all eigenvectors of non-zero eigenvalue are orthogonal to $\mathbf{1}$.)

Then $L_1(\tilde{Y}) = 2n\sum_{i=1}^{n}(b_{ii} - \sum_{k=1}^{n}\tilde{\lambda_k} v_{ik}^2)^2 + 2(\sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\tilde{\lambda_i})^2 + 4\sum_{i=1}^{n}(\lambda_i - \tilde{\lambda_i})^2$. Thus, choose top-$k$ $\tilde{\lambda_i} = \lambda_i$ might not be optimal.

Remarks:

Let $r = rank(B), s = Number\ of\ positive\ eigenvalues\ of\ B,$

1.If $k = r = s$,we could have perfectly embedding in $\mathbb{R}^k$,that is, there exists $\hat{Y} \in \mathbb{R}^k$,such that $d_{ij}^2 = \|\hat{Y}_i - \hat{Y}_j\|^2,\ B = \hat{Y}^T\hat{Y}$. Hence, both $L_0(Y)$ and $L_1(Y)$ has same minimization by $L_0(\hat{Y}) = L_1(\hat{Y}) = 0$.

2.If $k < r$ or $r > s$, $L_0(Y)$ and $L_1(Y)$ may be different, an intuitive idea comes from (3) : minimizing$\|R\|_F^2$,as we know,requires us to select engenvalues as large as we can,however,minimizing $(tr(R))^2$ requires to select middle engenvalues to balance the effect of negative ones. A trade-off here yields the result.

3. $r > s$ may happen if D does not come from Euclidean distance(e.g. spherical distance),or D comes from Euclidean distance but affected by some noise.

In fact, (2) is of the so-called form,sum of squares(SOS),which has tractable method to solve. Some numerical simulations could be found in [Tsang].Also,an effective method,SOS relaxation,to solve(2) in [Nie] is also recommended.