proof

@闫朝阳 2022.03.28

定义输入矩阵为X_{m*n}，m为维度，n为样本；

对X_{m*n}进行KernelPCA分析，即等价于，将X_{m*n}映射到高维空间得到$F_{d*n}(d>>m)$对其进行PCA分析。

记：$F=\varphi(X)$为由X到F的映射函数。而往往，$\varphi()$和F不可知。因此，对于高维空间的F_{d*n}的样本间内积$F\cdot F^T$难以计算，此时，设定义核函数$k(x,y)=\varphi(x)*\varphi(y)'$，表示$X$映射到高维空间的内积，可直接由$X$应用核函数等价。

即，核函数$k(X)$确定核矩阵K_{n*n}，等价于$F^T\cdot F$。

此时，对F_{d*n}进行PCA，
假设F_{d*n}已中心化，
则协方差矩阵$D_{d*d}=\frac{1}{n}F_{d*n}\cdot F^T_{n*d}$ $[\mathbf{1}]$，D_{d*d}为实对称矩阵，有性质：

> D_{d*d}可被对角化，且有$U^T\cdot D_{d*d}\cdot U=\Lambda$，其中$\Lambda$为对角元素为$D_{d*d}$特征值组成的对角矩阵；$U_{d*d}$每一列为特征值对应的特征向量；
> 其中，D_{d*d}求特征值特征向量可以表示为：$D\cdot U=\lambda U$ $[\mathbf{2}]$。

我们知道，对F_{d*n}进行PCA，即求协方差矩阵D_{d*d}的前k个特征向量按行组成的矩阵P_{k*d}使得
$P_{k*d}\cdot F_{d*n}=\dot{\boldsymbol{F}}_{k*n}$ $[\mathbf{3}]$；
此时$\dot{\boldsymbol{F}}_{k*n}$是我们希望的输出；
由PCA可知 (这里对PCA不做讨论)，
即，P_{k*d}等于上述前k特征值对应的特征向量组成的矩阵，即矩阵U的前k列的转置矩阵，U^T_{k*d}，
即，$U^T_{k*d}\cdot F_{d*n}=\dot{\boldsymbol{F}}_{k*n}$ $[\mathbf{4}]$

此时，联立$[\mathbf{1,2}]$式：
$D\cdot U=\lambda U$，
即，$\frac{1}{n}F_{d*n}\cdot F^T_{n*d}\cdot U=\lambda U$，
两边左乘F^T矩阵，
则，$\frac{1}{n}F^T\cdot F\cdot F^T\cdot U=\lambda F^T\cdot U$，
将$F^T\cdot F$视作整体，
即，$(F^T\cdot F)\cdot(F^T\cdot U)=n\lambda(F^T\cdot U)$，
显然，$F^T\cdot F$特征值为$n\lambda$，特征向量为$F^T\cdot U$ $[\mathbf{5}]$，
而根据$[\mathbf{4}]$式，我们希望的输出是：
$\dot{\boldsymbol{F}}_{k*n}=U^T_{k*d}\cdot F_{d*n}=(F^T_{n*d}\cdot U_{d*k})^T$，
即，为$[\mathbf{5}]$式中$F^T\cdot F$的前k个特征值对应的特征向量。
因此，我们只需对$F^T\cdot F$进行特征分解即可。
而$F^T\cdot F$，正是我们已知的核矩阵K_{n*n}。

解答2 与KPCA

当前方案中，
workflow先对输入矩阵X_{m*n}进行距离度量获得dissimilaity matrix M_{n*n}；
然后workflow对M_{n*n}标准化后得到$\bar{M}_{n*n}$，我们假设由X到$\bar{M}$映射函数为$\phi(x)$；
workflow然后对$\bar{M}$进行PCA分析并获取其前k个主成分，

也就是，获取了$\bar{M}_{n*n}$的协方差矩阵T_{n*n}的前k个特征值对应的特征向量矩阵；
我们形式化为：
$T_{n*n} = \frac{1}{n}\bar{M}_{n*n} \cdot \bar{M}_{n*n}^T$ **[6]**，
将$n \cdot T_{n*n} = \bar{M}_{n*n} \cdot \bar{M}_{n*n}^T$可等价于上述核矩阵$K_{n*n}$，
则，以上做法形式上等价于对X进行KPCA。
另外，$n \cdot T_{n*n}$核矩阵为协方差矩阵，因此一定为半正定矩阵，(reference)，满足核函数mercer定理；因此**核函数形式上等价**；

> Mercer 定理：任何半正定的函数都可以作为核函数。所谓半正定的函数f(xi,xj)，是指拥有训练数据集合（x1,x2,...xn)，我们定义一个矩阵的元素aij = f(xi,xj)，这个矩阵式n*n的，如果这个矩阵是半正定的，那么f(xi,xj)就称为半正定的函数。

但， 这里$\bar{M}_{n*n}$已知，由X经$\phi(x)$映射，并非向高维空间映射，因此，并**不严格意义等价。**

疑问
能否不经过dissimilarity matrix度量，而是直接获取到的就是核矩阵K，此时直接对核矩阵K进行特征分解即严格满足。或者说，我获取到的dissimilarity matrix不做协方差矩阵运算，而直接进行特征分解？

解答1 与PCoA

PCoA中，对dissimilaity matrix构建离差矩阵（实对称矩阵），该矩阵同样具有性质:

> $U^T \cdot D_{d*d} \cdot U = \Lambda$，其中$\Lambda$为对角元素为$D_{d*d}$特征值组成的对角矩阵；$U_{d*d}$每一列为特征值对应的特征向量； 即， $D_{d*d} = U \cdot \Lambda \cdot U^T$

进而，$D_{d*d} = U \cdot \Lambda \cdot U^T = U \cdot \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} \cdot U^T$，

$C = \Lambda^{\frac{1}{2}} \cdot U^T$即为坐标矩阵； 即取前$k$个特征值对应的特征向量按行组成，并乘各自特征值的平凡根，即为PCoA输出。

当前方案中，
workflow先对输入矩阵X_{m*n}进行距离度量获得dissimilaity matrix M_{n*n}；
然后计算协方差矩阵T_{n*n}，等价于PCoA中获得离差矩阵；
然后对T_{n*n}进行特征分解获取前k个特征值对应特征向量，相当于$C = U^T$而没有考虑特征值。
此外，有一点值得注意，workflow中dissimilaity matrix送入PCA中之前经过了一次数据标准化，导致原本dissimilaity matrix的数据结构性被破坏了。也就不满足PCoA的输入形式。因此，从这两个角度理解下来，**workflow做法都不等价于PCoA。**