Unsupervised Learning (2018 Fall)
Homework #3

Xinyuan Cao - `xc2461`

Nov 25 2018

Problem 1

1. **Nearest neighbor preserving embeddings by Indyk and Naor.**
This paper introduces nearest neighbor preserving embeddings, which means randomized embeddings between two metric spaces that maintain the approximate nearest neighbors. This can be very time efficient when given a new data, and to quickly find an approximate neighbor. In the literature we know that this task is closely related to low-distortion embeddings. And more recently, we've found that the embedding must be oblivious to X, and that the embedding does not need to preserve all interpoint distances. Here we define a good NN preserving embedding as follows:
Let $(Y, d_Y), (Z, d_Z)$ be metric spaces and $X \subset Y$. We say that a distribution over mappings $f : Y \to Z$ is a nearest neighbor preserving embedding (NN-preserving) with distortion $D \geq 1$ and probability of correctness $P \in [0, 1]$ if for every $c \geq 1$ and any $q \in Y$, with probability at least P, if $x \in X$ is such that $f(x)$ is a c-approximate nn of $f(q)$ in $f(X)$, then x is a $D \cdot d$ approximate nn of q in X. It is easy to see that JL lemma is an example of NN-preserving embedding. So here we focus on NN-preserving embeddings into low-dimensional spaces. And prove that such embedding exist for the following subset of l_x^d: doubling sets and sets with small aspect ratio and small $\gamma-$dimension. This algorithm has several applications to efficient approximate nearest neighbor problems. The first application combines NN-preserving embeddings with efficient $(1+\epsilon)$-approximate nearest neighbor data structures in l_2^k using $O(|x|/\epsilon^k)$ space and $O(k \log(|X|/\epsilon))$ query time. The second application is approximating NN where the data set contains objects that are more complex than points. So the approach can have simplicity preservation and modularity.

So here we introduce some basic concepts. Let (X, d_X) be a metric space. $B_X(x, r) = \{y \in X : d_X(x, y) < r\}$. We define the **doubling constant** of X (denoted as λ_X) is the least integer $\lambda \geq 1$ s.t. for every $x \in X$ and $r > 0$ there is $S \subset X$ with $|S| \leq \lambda$ s.t.

$$B_X(s, 2r) \subset \bigcup_{s \in S} B_X(s, r)$$

Fix $N \in Z$, $g = (g_1, \cdots, g_N)$ is a standard Gaussian vector in $\mathbb{R}^N$. Given $X \subset l_2^N$, denote the **parameter**

$$E_X = \mathbb{E} sup_{x\in X}| < x, g > |$$

We observe that for every bounded $X \subset l_2^N$:

$$E_X = O(diam(X)\sqrt{\log \lambda_X})$$

Given a metric space (X, d_X) set $\gamma_2(X) = \inf \sup_{x\in X} \sum_{s=0}^{\infty} 2^{s/2} d_X(x, A_s)$, where inf is taken over $A_S \subset X$ with $|A_S| < 2^{2^S}$. Define γ **dimension** of X to be:

$$\gamma dim(X) = [\frac{\gamma_2(X)}{diam(X)}]^2$$

Now we consider the case of Euclidean space with low $\gamma-$dimension. Let $(X, d_X), (Y, d_Y)$ be metric spaces, and $\delta, D > 0$. A mapping $f : X \to Y$ is said to be D **bi-Lipschitz with solution** δ if there is a scaling factor $C > 0$ s.t.

$$\forall a, b \in X, d_X(a, b) > \delta \Rightarrow Cd_X(a, b) \le d_Y(f(a), f(b)) \le CDd_X(a, b)$$

S^{d-1} denotes the unit Euclidean sphere centered at the origin. Using Gordon's theorem we know that fix $X \subset S^{d-1}$ and $\epsilon \in (0, 1)$. There exists an integer $k = O(\frac{E_X^2}{\epsilon^2})$ and a linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ s.t. $\forall x \in X$, $1 - \epsilon \le ||Tx||_2 \le 1 + \epsilon$. Now let X be a symmetric random variable s.t. $\mathbb{E}X^2 = 1$. Assume X is sub-Gaussian. Fix a unit vector $a \in S^{n-1}$. Let X_i be iid copies of X and denote $U = \sum_{j=1}^{n} a_j X_j$. Then $\mathbb{E}U^2 = 1$, and for every $0 \le t \le \frac{1}{8c}$ we have

$$\begin{aligned}
\mathbb{E}e^{tU^2} =& \mathbb{E}_U \mathbb{E}_g exp(\sqrt{2t} gU) = \mathbb{E}_g(\prod_{j=1}^{n} \mathbb{E}_X exp(\sqrt{2t} a_j gX)) \\
\le& \mathbb{E}_g[\prod_{j=1}^{n} exp(2cta_j^2 g^2)] = \mathbb{E}_g e^{2ctg^2} = \frac{1}{\sqrt{1-4ct}} \le \sqrt{2}
\end{aligned}$$

So we can see that for every $-\frac{1}{8c} \le t \le \frac{1}{8c}$ we have

$$\begin{aligned}
\mathbb{E}exp(tU^2) =& \sum_{m=0}^{\infty} \frac{t^m \mathbb{E}U^{2m}}{m!} \le 1 + t + \sum_{m=2}^{\infty} \frac{(8c|t|)^m (1/8c)^m \mathbb{E}Y^{2m}}{m!} \\
\le& 1 + t + (8ct)^2 \mathbb{E}exp(U^2/8c) \le 1 + t + 100c^2t^2 \le exp(t + 100c^2t^2)
\end{aligned}$$

So we can prove that for every $0 < \epsilon < 25c$ we have

$$Pr[|\frac{1}{k}\sum_{i=1}^{k} U_i^2 - 1| \ge \epsilon] \le 2exp(-k\epsilon^2/(400c^2))$$

And we can get a corollary that fix $\epsilon, \delta > 0$, $X \subset \mathbb{R}^d$. There exists an integer $k = O(\frac{E_X^2}{\delta^2\epsilon^2})$ s.t. X embeds $1+\epsilon$ bi-Lipschitzly in $\mathbb{R}^k$ with resolution δ. THe embedding extends to a linear mapping defined on all of $\mathbb{R}^d$. We can make this cor scale invariant by normalizing by diam(X). Here we can get a $1+\epsilon$ bi-Lipschitz embedding with resolution δ diam(X), where $k = O(\frac{\gamma dim(X)}{\delta^2\epsilon^2})$.

For the case of euclidean doubling spaces. We can prove the following theorem: For $X \subset \mathbb{R}^d$, $\epsilon \in (0,1)$ and $\delta \in (0, 0.5)$ there exists $k = (O(\frac{\log(2/\epsilon)}{\epsilon^2}\log(1/\delta)\log\lambda_X))$ such that for every $x_0 \in X$ with probability at least $1-\delta$,

$$d(Gx_0, G(X - \{x_0\})) \leq (1+\epsilon)d(x_0, X - \{x_0\})$$

For every $x \in X$ with $||x_0 - x||_2 > (1+2\epsilon)d(x_0, X - \{x_0\})$ satisfies $||Gx_0 - Gx|| > (1+\epsilon)d(x_0, X - \{x_0\})$.

2. **Optimality of the Johnson-Lindenstrauss lemma by Larseni and Nelson.**

This paper proves a lower bound for JL Lemma on the dimension m of the image space. Through literature review, a lower bound is obtained for only *linear* mappings and for a selected range of error ϵ or sufficiently large size n. The main contribution of this paper is giving a lower bound of m and thus proving the optimality of the JL Lemma. For thm.2, we get that: for any integers $n, d \geq 2$ and $\epsilon \in (\log^{0.5001} n/\sqrt{\min(n,d)}, 1)$, there exists a set of points $X \subset \mathbb{R}^d$ of size n, s.t. any map $f : X \to \mathbb{R}^m$ providing the guarantee of JL must have $m = \Omega(\epsilon^{-2}\log n)$

We proved JL Lemma through ramdom projection theories and indeed, all known proofs of the JL lemma proceed by instantiating distributions $D_{\epsilon,\delta}$ satisfying the guarantee of the below DJL lemma. (Actually JL lemma is a colloary of DJL Lemma when $\delta < 1/\binom{n}{2}$)

For any integer $d \geq 1$ and any $0 < \epsilon, \delta < \frac{1}{2}$, there exists a distribution $D_{\epsilon,\delta}$ over $m \times d$ real matrices for some $m \lesssim \epsilon^{-2}\log(\frac{1}{\delta})$ s.t.

$$\forall u \in \mathbb{R}^d, \underset{\Pi \sim D_{\epsilon,\delta}}{\mathbb{P}}(|||\Pi u||_2 - ||u||_2| > \epsilon||u||_2) < \delta \tag{1}$$

We want to prove the result by constructing a large family $\mathcal{P} = \{P_1, P_2, \cdots\}$ of very different sets of n points in $\mathbb{R}^d$. Through the whole proof, we assume all point sets in $\mathcal{P}$ can be embedded into $\mathbb{R}^m$ while preserving all pairwise distances to within $(1+\epsilon)$. If the data sets P_i are very different between each other, we know we cannot embed them into a very low-dimensional space. And the main idea is to construct special sets to argue that m should at least at some level to let this assumption hold.

First assume $d = \frac{n}{\log(1/\epsilon)}$ and $\epsilon \in (\log^{0.5001} n/\sqrt{d}, 1)$. For any set $S \subset [d]$ of $k = -\epsilon^{-2}/256$ indices, define $y_S := \sum_{j \in S}\frac{e_j}{\sqrt{k}}$, where e_i's denote the standard unit vectors in $\mathbb{R}^d$. So that $\langle e_j, y_S\rangle$ equals to 0 if $j \notin S$ and 16ϵ if $j \in S$. We then have $T = \{0, e_1, \cdots, e_d, y_S\}$, where $\forall u, v \in T, \langle f(u), f(v)\rangle \in (\langle u, v\rangle - 4\epsilon, \langle u, v\rangle + 4\epsilon)$.

Secondly, to extend d to any range, we give an encoding argument. Based on the assumption that the embedding preserves pairwise distances to with $(1+\epsilon)$, we can take any point set $T \in \mathcal{P}$ and encode it into a bit sr=tring of length O(nm). (The encoding guarantees that T can be uniquely recovered from the encoding.) The encoding algorithm thus effectively defines an injective mapping g from $\mathcal{P}$ to $\{0,1\}^{O(nm)}$. Since g is injective, we must have $|\mathcal{P}| \leq 2^{O(mn)}$. But $|\mathcal{P}| = \binom{d}{k}^Q = (\epsilon^2 n/\log(\frac{1}{\epsilon}))^{\Omega(\epsilon^{-2}n)}$ and we can conclude $m = \Omega(\epsilon^{-2}\log(\epsilon^2 n/\log(\frac{1}{\epsilon}))$. For $\epsilon > \frac{1}{n^{0.4999}}$, this is $m = \Omega(\epsilon^{-2}\log n)$.

Then it proposes the idea to reduce the length of the encoding to $O(mn)$. We can encode approximations $\hat{f}(e_j)$ by specifying indices into a covering C_2 of $(1+\epsilon)B_2^m$ by ϵB_2^m. Define a $d \times m$ matrix A having $\hat{f}(e_j) = c_2 f(e_j)$ as rows. The by utilizing the matrix and the structure of $\mathcal{P}$, we can get down to just $O(m)$ bits. Denote W as the subspace of $\mathbb{R}^d$ spanned by the columns of A and define $U := B_\infty^d \cap W$ as the convex body. Now let C_∞ be a minimum cardinality covering of $(22\epsilon)U$ by translated copies of ϵU. By the following lemma and corollary, we have $|C_\infty| \leq 2^{m\log 45}$. For lemma 2 in the paper: Let E be an m-dimensional normed space and let B_E denote its unit ball. For any $0 < \epsilon < 1$, one can cover B_E using at most $2^{m\log(1+2/\epsilon)}$ translated copies of ϵB_E. And we have the corollary: Let T be an origin symmetric convex body in $\mathbb{R}^m$. For any $0 < \epsilon < 1$, one can cover T using at most $2^{m\log(1+2/\epsilon)}$ translated copies of ϵT.

And finally we analyse the size of the encoding produced by the above procedure and derive a lower bound on m. Recall that the encodig procedure produces $O(mn)$ bits but $|\mathcal{P}| \geq (d/2k)^{kQ} \geq (d/2k)^{kn/2}$. Therefore $m = \Omega(\epsilon^{-2}\log(\epsilon^2 n/\log(1/\epsilon)))$. With the assumption $\epsilon > \log^{0.5001} n/\sqrt{d}$ and this can be simplified to $m = \omega(\epsilon^{-2}\log(\epsilon^2 n))$.

To note that in the previous steps we utilize some lemma to get our conclusion.

For every e_j and y_{Sl} in T, we have

$$|\langle \hat{f}(e_j), \hat{f}(y_{Sl})\rangle - \langle e_j, y_{Sl}\rangle| \leq 6\epsilon$$

Finally we make some handling toward other values of d by differently constructing the point sets in $\mathcal{P}$ and discuss the cases in which $d > n/\log(1/\epsilon)$ and in which $d > n/\log(1/\epsilon)$ and $\epsilon \in (\log^{0.5001} n/\sqrt{d}, 1)$ through proof by contradiction.

Problem 2

For any $u, v \in S$, denote $x = u - v$. We want show that for any $0 < \epsilon < 1/2$ there exists a linear map $f : \mathbb{R}^D \to \mathbb{R}^d$ s.t.

$$(1-\epsilon)||u-v|| \leq ||f(u) - f(v)|| \leq (1+\epsilon)||u-v||$$

For a linear map f, we have $f(u) - f(v) = f(u-v) = f(x)$, and $f(x) = ||x||f(\frac{x}{||x||})$. So it is suffice to show that for any x in the unit ball of S,

$$(1-\epsilon)||x|| \leq ||f(x)|| \leq (1+\epsilon)||x|| \tag{2}$$

Let $\epsilon_0 = \epsilon/3$, so we have $0 < \epsilon_0 < 1/6 < 1/2$. Let $r = \frac{\epsilon}{3+\epsilon}$, so we have $\frac{r}{1-r} = \epsilon/3$. Denote S_0 as the unit ball of S. $S_0 \subset S \subset \mathbb{R}^D$, and S_0 is close and bounded, so it is compact. Hence we can find a finite $r-$cover on S_0, denote this cover as C. Because C is finite, we can apply JL-Lemma to C. That is, for ϵ_0, there exists a linear map $f : C \to \mathbb{R}^d$ (with $d = O(k/\epsilon^2)$), such that for all $u, v \in C$,

$$(1-\epsilon_0)||u-v|| \leq ||f(u) - f(v)|| \leq (1+\epsilon_0)||u-v|| \tag{3}$$

Because $\mathbb{R}^D$ and $\mathbb{R}^d$ are finite dimensional vector spaces with defined basis, we know f can be represented by a $d \times D$ matrix A. So consider $\phi : S_0 \to \mathbb{R}^d$ such that $\phi(x) = Ax$ for $x \in S_0$. ϕ is a linear map. So $\phi(0) = \phi(u-u) = \phi(u) - \phi(u) = 0$. Let $v = 0$ in equation (2) and we get for all $u \in C$,

$$1 - \epsilon_0 = (1-\epsilon_0)||u|| \leq ||f(u)|| = ||\phi(u)|| \leq (1+\epsilon_0)||u|| = 1 + \epsilon_0 \tag{4}$$

So we know that

$$\max_{x \in C} ||\phi(x)|| \leq 1 + \epsilon_0 \tag{5}$$

$$\min_{x \in C} ||\phi(x)|| \geq 1 - \epsilon_0 \tag{6}$$

Then for any x in S_0, since C is an $r-$cover of S_0, there exists a $x_c \in C$ such that $x = x_c + x_r$ and $||x_r|| \leq r$.

$$||\phi(x)|| = ||\phi(x_c) + \phi(x_r)|| \leq ||\phi(x_c)|| + ||\phi(x_r)|| \tag{7}$$

Recall Question 4(iv) of HW1, we know that

$$\max_{x \in S_0} ||\phi(x)|| \leq \frac{\max_{x \in C} ||\phi(x)||}{1-r} \leq \frac{1+\epsilon_0}{1-r}$$

So we know that

$$\max_{||x_\delta||=\delta} ||\phi(x_\delta)|| = \delta \max_{||x_\delta||=\delta} ||\phi(x_\delta/\delta)|| = \delta \max_{||x||=1} ||\phi(x)|| \leq \delta \frac{1+\epsilon_0}{1-r}$$

So we have

$$\max_{||x_r|| \leq r} ||\phi(x_r)|| \leq \frac{r(1+\epsilon_0)}{1-r} \tag{8}$$

Therefore we can derive from equation(4)(6)(7) that

$$\begin{aligned}
\max_{x \in S_0} ||\phi(x)|| \leq & \max_{x_c \in C} ||\phi(x_c)|| + \max_{||x_r|| \leq r} ||\phi(x_r)|| \\
\leq & (1+\epsilon_0) + \frac{r(1+\epsilon_0)}{1-r} \\
= & \frac{1+\epsilon/3}{1-\frac{\epsilon}{3+\epsilon}} \\
= & (\epsilon+3)^2/9 \\
\leq & 1+\epsilon
\end{aligned}$$

From here we have proved the right part of (1). For the other part, we need to estimate the lower bound of $||\phi(x)||$.

$$||\phi(x)|| = ||\phi(x_c) + \phi(x_r)|| \geq \phi(x_c) - \phi(x_r) \tag{9}$$

We need to estimate the minimum of $||\phi(x)||$. From equation (5)(7)(8)

$$\begin{aligned}
\min_{x \in S_0} ||\phi(x)|| \geq & \min_{x_c \in C} ||\phi(x_c)|| - \max_{||x_r|| \leq r} ||\phi(x_r)|| \\
\geq & 1 - \epsilon_0 - \frac{r(1+\epsilon_0)}{1-r} \\
= & 1 - \epsilon/3 - (1+\epsilon/3)\epsilon/3 \\
= & 1 - 2\epsilon/3 - \epsilon^2/9 \\
\geq & 1 - 2\epsilon/3 - \epsilon/3 \\
= & 1 - \epsilon
\end{aligned}$$

So now we have proved that for any x in the unit ball, equation (1) is correct.
Finally we estimate d. Recall Question 4(iii) of HW1, we know that $r-$cover number $\geq (\frac{1}{r})^k$, so from JL-Lemma we know that

$$d = \Omega(\frac{\log n}{\epsilon_0^2}) = \Omega(\frac{k \log \frac{1}{r}}{\epsilon_0^2}) = \Omega(\frac{k \log(1+\frac{3}{\epsilon})}{\epsilon^2/9}) = \Omega(k/\epsilon^2)$$

Therefore we have proved the JL-type result in a fixed but unknown k-dimensional affine space. #
- Reference: Discuss with Yuze Zhou.

Problem 3

1. Since P is the projection onto the orthogonal complement of $span(\mathbf{1}_v)$, we have $P = I - \frac{1}{n}\mathbf{1}_v\mathbf{1}_v^T$, and $P\mathbf{1}_S = \mathbf{1}_S - \frac{|S|}{n}\mathbf{1}_v$. Then we have $||P\mathbf{1}_s||_2^2 = |S| + \frac{|S|^2}{n} - 2\frac{|S|^2}{n} = |S| - \frac{|S|^2}{n}$. Since $|S| \leq \frac{|V|}{2}$ we have

$$\frac{||P\mathbf{1}_s||_2^2}{||\mathbf{1}_S||_2^2} = 1 - \frac{|S|}{n} \geq 1 - \frac{1}{2} = \frac{1}{2}$$

Recall the properties we've derived in the class and in homework 1 and 2, we know L has an eigenpair $(0, \mathbf{1})$, so $I(L) \subset span(\mathbf{1}_v)^\perp$, here $I(L)$ denotes the image of L. Thus $PL = L = LP$. So we have

$$\phi = \min_{x \in \{\mathbf{1}_S : |S| \leq \frac{|V|}{2}\}} \frac{1}{d}\frac{x^T L x}{||x||_2^2} = \min_{x \in \{\mathbf{1}_S : |S| \leq \frac{|V|}{2}\}} \frac{1}{d}\frac{(Px)^T LPx}{||Px||_2^2}\frac{||Px||_2^2}{||x||_2^2}$$

$\forall x \in span(\mathbf{1}_v)^\perp$, $R(x) \geq \lambda_2$,

$$\frac{1}{d}\frac{(Px)^T LPx}{||Px||_2^2}\frac{||Px||_2^2}{||x||_2^2} \geq \frac{1}{d}\lambda_2\frac{||Px||_2^2}{||x||_2^2} \geq \frac{\lambda_2}{2d}$$

Minimize both side on x, and so we get $\phi \geq \frac{\lambda_2}{2d}$.

2. With the assumption that there exist $\mathcal{D}$ such that $\mathbb{P}_{k\sim D}\left(\phi(S_k) \leq \sqrt{\frac{2\lambda_2}{d}}\right) > 0$,

$$0 < \mathbb{P}_{k\sim D}\left(\phi(S_k) \leq \sqrt{\frac{2\lambda_2}{d}}\right) = \mathbb{E}_{k\sim D}\left(\mathbb{I}_{\left\{\phi(S_k) \leq \sqrt{\frac{2\lambda_2}{d}}\right\}}\right)$$

So there exist at least one k such that satisfies $\phi(S_k) \leq \sqrt{\frac{2\lambda_2}{d}}$. So we have our algorithm as follows.

Algorithm 1 edge expansion algorithm

function EDGEEXPANSION(G(V,E)) ▷ $V = \{v_1, \cdots, v_n\}$
    Compute x as the minimizer of modified relaxation problem
    Resort all the nodes such that $x_1 \leq x_2 \leq \cdots \leq x_n$, $T = S_1 = \phi(\{v_1\})$, $Out = \{v_1\}$
    for i = 2:(n - 1) **do**
        Construct $S_i = \{v_1, \cdots, v_i\}$
        if $\phi(S_i) \leq T$ **then**
            $T = \phi(S_i)$
            if $|S_i| \leq \frac{n}{2}$ **then**
                $Out = S_i$
            else
                $Out = V \setminus S_i$
            end if
        end if
    end for
    T and Out
end function

3. Since L has only one eigenvalue that equals to 0, and all other eigenvalues are greater than 0. We obtain the min v from

$$R(v) = \min_{x \in span(\mathbf{1}_v)^\perp} \frac{x^T L x}{||x||_2^2}$$

Let v' be the eigenvector corresponding to λ_2, so we have

$$R(v') = \frac{\lambda_2 ||v'||_2^2}{||v'||_2^2} = \lambda_2 \geq R(v)$$

By Rayleigh quotient and SVD, we know v' is also a solver to the minimization problem. If $u = \alpha(v + \beta \mathbf{1}_v)$, so

$$R(u) = \frac{\alpha^2 (v + \beta \mathbf{1}_v)^T L (v + \beta \mathbf{1}_v)}{\alpha^2 ||v + \beta \mathbf{1}_v||_2^2} = \frac{v^T L v}{||v + \beta \mathbf{1}||_2^2} = \frac{v^T L v}{||v||_2^2 + \beta^2 ||\mathbf{1}_v||_2^2} \leq \frac{v^T L v}{||v||_2^2} = R(v)$$

This holds since v and $\mathbf{1}_v$ are orthogonal and $\mathbf{1}_v$ is the eigenvector of L corresponding to eigenvalue 0.

4. If $(i,j) \in E$ and $i < j$, we have

$$\mathop{\mathbb{P}}_{k \sim D}(i \leq k < j) = \mathop{\mathbb{P}}_{k \sim D}((i,j) \in E(S_k, \bar{S}_k)) = \sum_{k=i}^{j-1} |u_{k+1}^2 - u_k^2|$$

If $u_i \geq 0$ or $u_j \leq 0$, $\mathop{\mathbb{P}}_{k \sim D}(i \leq k < j) = |u_j^2 - u_i^2| = |u_i - u_j||u_i + u_j| \leq |u_i - u_j|(|u_i| + |u_j|)$.

If $u_i < 0 < u_j$, find index t such that $u_t \le 0$ and $u_{t+1} \ge 0$, (actually $t = \lfloor \frac{|V|}{2} \rfloor - 1$) then $\mathbb{P}_{k\sim D}(i \le k < j) = u_i^2 - u_t^2 + u_j^2 - u_{t+1}^2 \le u_i^2 + u_j^2 \le (|u_i| + |u_j|)^2 = |u_i - u_j|(|u_i| + |u_j|)$

Moreover, we can compute the expectation of the cardinality of the cut as

$$\mathop{\mathbb{E}}_{k\sim D}(|E(S_k, \bar{S}_k)|) = \mathop{\mathbb{E}}_{k\sim D}\left(\sum_{(i,j)\in E, i<j} (\mathbb{I}_{(i,j)\in E(S_k, \bar{S}_k)})\right) = \sum_{(i,j)\in E, i<j} \mathop{\mathbb{P}}_{k\sim D}(i \le k < j)$$

Therefore $\mathop{\mathbb{E}}_{k\sim D}(|E(S_k, \bar{S}_k)|) \le \sum_{(i,j)\in E, i<j} |u_i - u_j|(|u_i| + |u_j|)$

5. If $i < \frac{|V|}{2}$, i would be assigned to S_k if and only if $i \le k < \frac{|V|}{2}$. Since $u_{\lfloor \frac{|V|}{2} \rfloor - 1} = 0$

$$P(i \in S_k) = \sum_{j=i}^{\lfloor \frac{|V|}{2} \rfloor - 1} |u_{j+1}^2 - u_j^2| = \sum_{j=i}^{\lfloor \frac{|V|}{2} \rfloor} (u_j^2 - u_{j+1}^2) = u_i^2 - u_{\lfloor \frac{n}{2} \rfloor}^2 = u_i^2$$

For $i \ge \frac{|V|}{2}$, i would be assigned to S_k if and only if $i > k \ge \frac{|V|}{2}$,

$$P(i \in S_k) = \sum_{j=\lfloor \frac{|V|}{2} \rfloor}^{i-1} |u_{j+1}^2 - u_j^2| = \sum_{j=\lfloor \frac{|V|}{2} \rfloor}^{i-1} (u_{j+1}^2 - u_j^2) = u_i^2$$

Thus

$$\mathop{\mathbb{E}}_{k\sim D}(vol(S_k)) = \mathop{\mathbb{E}}_{k\sim D}(d \sum_{i=1}^{n} \mathbb{I}_{i\in S_k}) = d \sum_{i=1}^{n} P(i \in S_k) = d \sum_{i=1}^{n} u_i^2$$

6. From (4), $\mathop{\mathbb{E}}_{k\sim D}(|E(S_k, \bar{S}_k)|) \le \sum_{(i,j)\in E, i<j} |u_i - u_j|(|u_i| + |u_j|)$, and Cauchy inequality tells us,

$$\mathop{\mathbb{E}}_{k\sim D}(|E(S_k, \bar{S}_k)|) \le \sqrt{\sum_{(i,j)\in E, i<j} |u_i - u_j|^2} \sqrt{\sum_{(i,j)\in E, i<j} (|u_i| + |u_j|)^2} = \sqrt{u^T L u} \sqrt{\sum_{(i,j)\in E, i<j} (|u_i| + |u_j|)^2}$$

Therefore

$$\frac{\mathop{\mathbb{E}}_{k\sim D}(|E(S_k, \bar{S}_k)|)}{\mathop{\mathbb{E}}_{k\sim D}(vol(S_k))} \le \frac{\sqrt{u^T L u}\sqrt{2d \sum_{i=1}^{n} u_i^2}}{d \sum_{i=1}^{n} u_i^2} = \sqrt{\frac{2u^T L u}{d \sum_{i=1}^{n} u_i^2}} = \sqrt{\frac{2R(u)}{d}}$$

7. Since probability is nonnegative, suppose $P\left(\frac{X}{Y} \le \frac{\mathbb{E}X}{\mathbb{E}Y}\right) = 0$. Then $\frac{X}{Y} > \frac{\mathbb{E}X}{\mathbb{E}Y}$ holds almost surely. Since X and Y are both positive random variables, $X\mathbb{E}Y > Y\mathbb{E}X$ also holds almost surely. Taking expectation for both sides, and we have $\mathbb{E}X\mathbb{E}Y > \mathbb{E}Y\mathbb{E}X$. This causes a contradiction. So we have

$$P\left(\frac{X}{Y} \le \frac{\mathbb{E}X}{\mathbb{E}Y}\right) > 0$$

8. Since $\phi(S_k) = \frac{|E(S_k, \bar{S}_k)|}{vol(S_k)}$, we have $\underset{k\sim D}{\mathbb{P}}\left(\phi(S_k) \leq \sqrt{\frac{2R(u)}{d}}\right) > 0$.

   From (3) we've prove $R(u) \leq R(v) = \lambda_2$, so we have

$$0 < \underset{k\sim D}{\mathbb{P}}\left(\phi(S_k) \leq \sqrt{\frac{2R(u)}{d}}\right) \leq \underset{k\sim D}{\mathbb{P}}\left(\phi(S_k) \leq \sqrt{\frac{2R(v)}{d}}\right) = \underset{k\sim D}{\mathbb{P}}\left(\phi(S_k) \leq \sqrt{\frac{2\lambda_2}{d}}\right)$$

- Reference: Discuss with Linyun He, Yuze Zhou.

Problem 4

1. From the definition of eigenvalue and eigenvector we have

$$\lambda_i v_i = XX^T v_i \tag{10}$$

$$\mu_i u_i = X^T X u_i \tag{11}$$

Lemma 1. *If A is a $m \times n$ matrix, B is a $n \times m$ matrix, $m \geq n$, then we have the non-zero eigenvalues of AB are the same as the non-zero eigenvalues of BA, and AB has $m - n$ more zero eigenvalues than BA.*

Proof.

$$\begin{bmatrix} I & B \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ A & AB \end{bmatrix} = \begin{bmatrix} BA & BAB \\ A & AB \end{bmatrix} = \begin{bmatrix} BA & 0 \\ A & 0 \end{bmatrix} \begin{bmatrix} I & B \\ 0 & I \end{bmatrix}$$

So we have $C_1 = \begin{bmatrix} 0 & 0 \\ A & AB \end{bmatrix}$ and $C_2 = \begin{bmatrix} BA & 0 \\ A & 0 \end{bmatrix}$ are similar. And thus have same eigenvalues. The eigenvalues of C_1 are the eigenvalues of AB and n zeros. The eigenvalues of C_2 are the eigenvalues of BA and m zeros. So we know that AB and BA have same non-zero eigenvalues, and AB has $m - n$ more zero eigenvalues than BA. □

From the lemma we know that XX^T and $X^T X$ has same non-zero eigenvalues. And the number of zero eigenvalues of XX^T is $D - n$ more than the number of zero eigenvalues of $X^T X$.

Because (μ_i, u_i) is the eigenvalue/vector pairs of $X^T X$, we have

$$\mu_i u_i = X^T X u_i$$

Left multiply the equation by X, and we have

$$\mu_i X u_i = XX^T X u_i$$

That is, $\mu_i(Xu_i) = XX^T(Xu_i)$. So we know that if $Xu_i \neq 0$, the pair (μ_i, Xu_i) is the eigenvalue/vector of XX^T.

For $Xu_i = 0$, from equation (10) we know that LHS = RHS = 0. Since $u_i \neq 0$, we know that $\mu_i = 0$. So $\lambda_i = 0$. Here the corresponding eigenvectors of eigenvalue 0 is the vectors in the null space of XX^T, which are orthogonal and span the null space of XX^T.

Therefore we can write (λ_i, v_i) as:

$$\begin{cases} (\mu_i, Xu_i) & \text{if } \mu_i \neq 0, \\ (0, v_t), \text{where } \{v_t\} \text{ are orthogonal and span the null space of } XX^T & \text{if } \mu_i = 0. \end{cases}$$

2. Let's do SVD decomposition of X, denote $X = VSU^T$, $V \in \mathbb{R}^{D\times D}$, $S \in \mathbb{R}^{D\times n}$, $U \in \mathbb{R}^{n\times n}$. So we have $XX^T = VS^2V^T$. So we know the diagonal of S is the square root of λ_i.

   We want to project the given data matrix X into the k dimensional PCA subspace span$(v_1, v_2, \cdots, v_k)$, where $v_1, \cdots, v_k$ is the corresponding eigenvectors of the k largest eigenvalues of XX^T. Denote $V_k = (v_1^T, \cdots, v_k^T)^T$, $S_k = diag(\lambda_1, \cdots, \lambda_k)$, $U_k = (u_1^T, \cdots, u_k^T)$. They are the largest k eigenvalues and corresponding eigenvectors.

$$\tilde{X} = V_k S_k U_k^T = \sum_{i=1}^{k} v_i \sqrt{\lambda_i} u_i^T = \sum_{i=1}^{k} \frac{Xu_i}{||Xu_i||} \sqrt{\lambda_i} u_i^T$$

   Since

$$\frac{Xu_i}{||Xu_i||} = \frac{Xu_i}{\sqrt{u_i^T X^T X u_i}} = \frac{Xu_i}{\sqrt{u_i^T \lambda_i u_i}} = \frac{Xu_i}{\sqrt{\lambda_i}}$$

   So we get

$$\tilde{X} = \sum_{i=1}^{k} X u_i u_i^T$$

3. For a new datapoint $x \in \mathbb{R}^D$, the k-dimensional PCA projection of x is

$$\tilde{x} = \sum_{i=1}^{k} v_i v_i^T x = \sum_{i=1}^{k} \frac{Xu_i}{\sqrt{\lambda_i}} (\frac{Xu_i}{\sqrt{\lambda_i}})^T x = \sum_{i=1}^{k} \frac{1}{\lambda_i} X u_i u_i^T X^T x$$

4. First consider a new data point x'. Let $\phi(X) = [\phi(x_1), \cdots, \phi(x_n)]$ be the coordinates of all data points in the feature space. Then the projection of x' onto the PCA subspace of the feature space is therefore $\sum_{i=1}^{k} \frac{\phi(X)\tilde{u}_i\tilde{u}_i^\tau \phi(X)^\tau}{\tilde{\lambda}_i}\phi(x')$, where $\tilde{u}_i$ are the top eigenvectors corresponding to the top eigenvalues $\tilde{\lambda}$ of $\phi(X)^\tau\phi(X)$. Without loss of generalities, let the SVD decomposition of $K = \phi(X)^\tau\phi(X)$ is $\tilde{V}\tilde{\Lambda}\tilde{U}$, where $\tilde{\Lambda}$ is a $D * n$ diagonal matrix containing the singular values $\sqrt{\tilde{\lambda}_i}$, and $\tilde{U} = [\tilde{u}_1, \cdots, \tilde{u}_n]$,$\tilde{V} = [\tilde{v}_1, \cdots, \tilde{v}_D]$. Let $K(X, x') = [K(x_1, x'), \cdots, K(x_n, x')]$ be the column vector of the inner product of the projection of x' onto the PCA subspace of the feature space could be further written as:

$$\begin{aligned}\sum_{i=1}^{k} \frac{\phi(X)\tilde{u}_i\tilde{u}_i^\tau \phi(X)^\tau}{\tilde{\lambda}_i}\phi(x') &= \sum_{i=1}^{k} \frac{\tilde{v}_i\sqrt{\tilde{\lambda}_i}\tilde{u}_i^\tau K(X, x')}{\sqrt{\lambda_i}} \\ &= \sum_{i=1}^{k} \frac{\tilde{u}_i^\tau K(X, x')}{\sqrt{\tilde{\lambda}_i}}\tilde{v}_i\end{aligned}$$

Since $\tilde{v}_1, \cdots, \tilde{v}_k$ are the orthonormal vectors that spann the k-dimensional PCA subspace in the feature space. We have the coordinates of the data onto the specific subspace with respect to $\tilde{v}_1, \cdots, \tilde{v}_k$ is $[\frac{\tilde{u}_1^\tau K(X,x')}{\sqrt{\tilde{\lambda}_1}}, \cdots, \frac{\tilde{u}_k^\tau K(X,x')}{\sqrt{\tilde{\lambda}_k}}]$.

Denote $\tilde{\Lambda}_k^{-1/2} = diag(1/\sqrt{\tilde{\lambda}_1}, \cdots, 1/\sqrt{\tilde{\lambda}_k})$ and $K = (<\phi(x_i), \phi(x_j)>)_{ij}$ be the inner product matrix of all old data points in the feature space, we can get the input data points' coordinates in the subspace is $\tilde{\Lambda}_k^{-1/2}\tilde{U}_k^\tau K$.

5. I use two concentric circles to show the kernalized-PCA result.

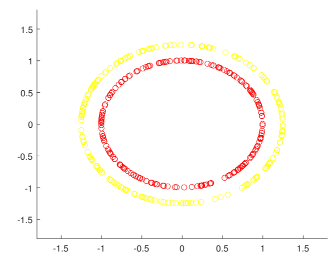


Figure 1: original data

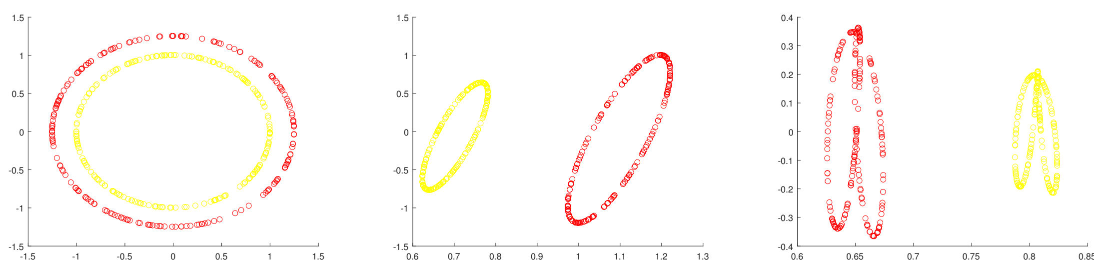


Figure 2: 2D PCA (linear, quadratic, rbf kernels)

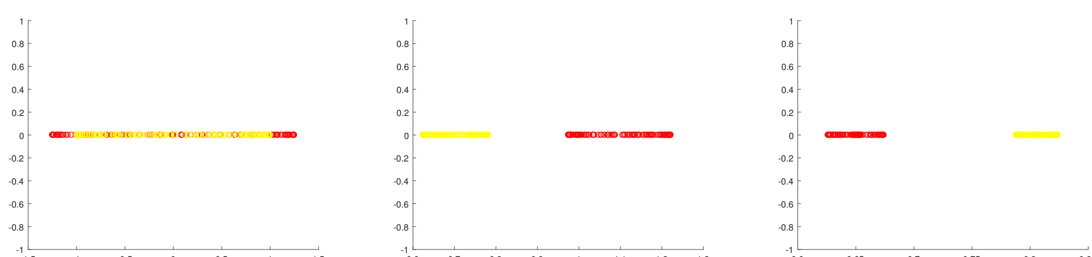


Figure 3: 1D PCA (linear, quadratic, rbf kernels)

We can see that 2D PCA has better result compared to 1D PCA. And rbf kernel has very good results for concentric circles. #

- Reference: https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca
- Discuss with Yuze Zhou

Problem 5

Here I use two methods to separate audio signals: fourth-order blind identification (FOBI) and fastICA.

Algorithm 2 Fourth-order Blind Identification

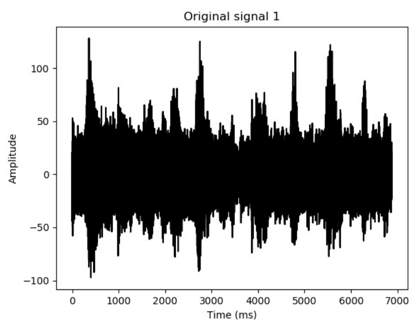
---

input: the mixed signals $X \in \mathbb{R}^{D\times T}$ after centering and whitening, number of signals k, number of mixtures D, time sampling number T
output: the separated signals $S \in \mathbb{R}^{k\times T}$

function FOURTH-ORDER BLIND IDENTIFICATION(X)
    $\Omega = E(XX^T||x||^2)$
    Y = eigenvectors of Ω ▷ find the eigenvectors of Ω
    return Y^TX
end function

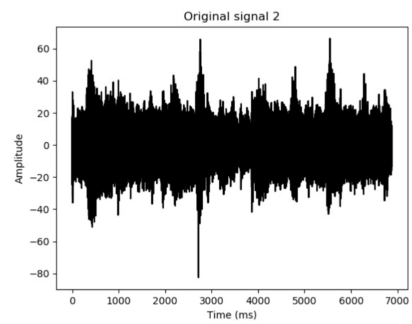
This method is very efficient and has very good efficient on dataset 1. I plot the original mixture signals and the separated signals as follows:

signal 1.jpg



(a) mixture signal 1

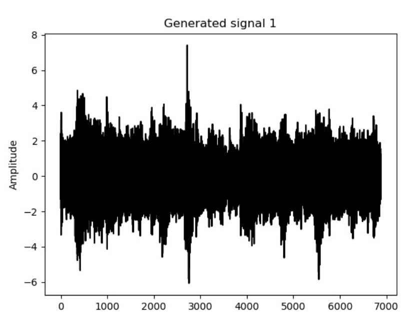
signal 2.jpg



(b) mixture signal 2

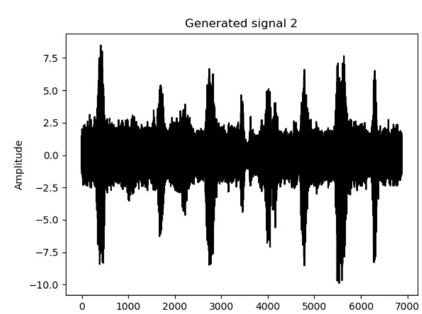
Figure 4: Mixture signals

signal 1.jpg



(a) separated signal 1

signal 2.jpg



(b) separated signal 2

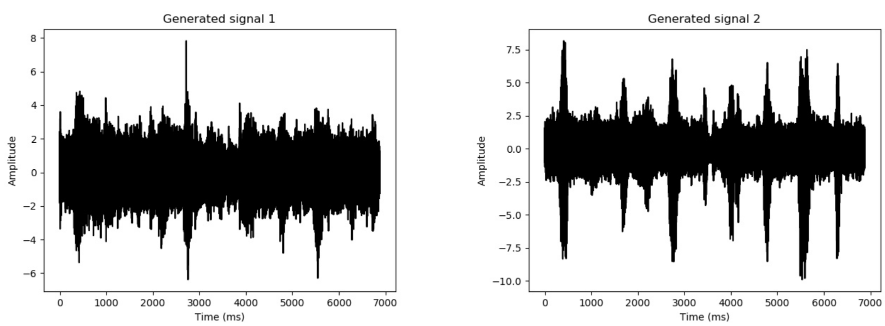
Figure 5: Separated signals using FOBI

So it is obvious that the result of this algorithm can separate the background music apart with the voice. This can also be identified by listening to the output file.
Secondly I follow the tutorial and implemented FastICA.

Algorithm 3 FastICA

input: the mixed signals $X \in \mathbb{R}^{D\times T}$ after centering and whitening, number of signals k, number of mixtures D, time sampling number, maximum iteration number T
output: the separated signals $S \in \mathbb{R}^{k\times T}$

function FOURTH-ORDER BLIND IDENTIFICATION(X)
  for i in 1 to k **do**
    Initialize W_p randomly
    while W_p changes **do**
      $W_p \leftarrow \frac{1}{T}Xg(w_p^T X)^T - \frac{1}{T}g'(w_p^T X)\mathbf{1}w_p$ ▷ Here we choose $g(u) = \tanh(u)$
      $W_p \leftarrow W_p - \sum_{j=1}^{i-1}(W_p^T w_j)w_j$
      $W_p \leftarrow \frac{W_p}{||W_p||}$
    end while
  end for
  return $W^T X$
end function



(a) separated signal 1 (b) separated signal 2

Figure 6: Separated signals using FastICA

It can seen and also heard that this method has similar performance as FOBI. But for dataset2 and 3, these two algorithms both cannot achieve good results.

- Reference

1. Independent Component Analysis A Tutorial by Hyvrinen and Oj
2. Source Separation Using Higher Order Moments by J.-F. Cardoso
3. https://www.cs.helsinki.fi/u/ahyvarin/papers/bookfinal_ICA.pdf