# COMS 4772 Fall 2016 Homework 2
Due Friday, October 28

**Instructions**:

- The usual homework policies (`http://www.cs.columbia.edu/~djhsu/coms4772-f16/about.html`) are, of course, in effect.
- Using this LaTeX template will be helpful for grading purposes.

**Problem 1** (25 points)**.** Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$ whose distribution is a mixture of $k$ spherical Gaussians:

$$\boldsymbol{X} \;\sim\; \pi_1 \, \mathrm{N}(\boldsymbol{\mu}_1, \sigma_1^2 \boldsymbol{I}) + \pi_2 \, \mathrm{N}(\boldsymbol{\mu}_2, \sigma_2^2 \boldsymbol{I}) + \cdots + \pi_k \, \mathrm{N}(\boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I}) \,.$$

For any set $C \subset \mathbb{R}^d$, define

$$\mathrm{cost}(C) \;:=\; \mathbb{E}\left[\min_{\boldsymbol{y} \in C} \|\boldsymbol{X} - \boldsymbol{y}\|_2^2\right] \,.$$

Let $M := \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k\}$. Prove that if $k < e^{d/2}$, then

$$\mathrm{cost}(M) \;\leq\; \frac{1}{1 - \frac{2\ln(k)}{d}} \cdot \min_{\substack{C \subset \mathbb{R}^d: \\ |C| \leq k}} \mathrm{cost}(C) \,.$$

*Solution.* $\square$

**Problem 2** (25 points)**.** Suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times d}$ each have rank $d$. Give unambiguous pseudocode for an algorithm that, when given $\boldsymbol{A}$ and $\boldsymbol{B}$ as inputs, finds all solutions $\boldsymbol{v} \in S^{d-1}$ satisfying

$$\exists \lambda \in \mathbb{R} \setminus \{0\} \text{ s.t. } \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{v} \;=\; \lambda \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{v}\,.$$

If there is an entire subspace of solutions, the algorithm just needs to return an orthonormal basis for this subspace. Your pseudocode can use things like SVD, Gram-Schmidt, etc. as black-box subroutines. Prove that the algorithm is correct.

*Solution.* □

**Problem 3** (25 points)**.** Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ be a data matrix whose rows are $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \in \mathbb{R}^d$. Let $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ be the matrix whose $(i,j)$-th entry is the squared Euclidean distance $D_{i,j} = \left\|\boldsymbol{a}_i - \boldsymbol{a}_j\right\|_2^2$. Suppose you are given the squared Euclidean distance matrix $\boldsymbol{D}$ as input, and you are asked to recover the set of original points $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$ up to some translation. *You do not have access to the original data matrix* $\boldsymbol{A}$*.*

(a) Let $\boldsymbol{s} \in \mathbb{R}^n$ be the vector whose $i$-th entry is $\|\boldsymbol{a}_i\|_2^2$. Prove that $\boldsymbol{D} = \boldsymbol{s}\mathbf{1}_n^\top - 2\boldsymbol{A}\boldsymbol{A}^\top + \mathbf{1}_n\boldsymbol{s}^\top$, where $\mathbf{1}_n \in \mathbb{R}^n$ is the all-ones vector.

(b) Let $\boldsymbol{\Pi} \in \mathbb{R}^{n \times n}$ be the orthogonal projector for the $(n-1)$-dimensional subspace

$$\left\{\boldsymbol{x} \in \mathbb{R}^n : \langle \mathbf{1}_n, \boldsymbol{x} \rangle = 0\right\} .$$

Prove that $-(1/2)\boldsymbol{\Pi}\boldsymbol{D}\boldsymbol{\Pi} = \boldsymbol{\Pi}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\Pi}$.

(c) Explain how to determine points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ from $\boldsymbol{D}$ such that:

- $D_{i,j} = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2$ for all $i, j \in [n]$; and
- $\sum_{i=1}^n \boldsymbol{x}_i = \mathbf{0}$.

(You may assume that you are told the original dimension $d$.)

(d) *Optional.* Suppose the matrix $\boldsymbol{D}$ is corrupted (say, because your distance measuring device is imperfect), so the entries no longer correspond to the squared Euclidean distances between the $\boldsymbol{a}_i$. Explain how to determine points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^n$ (yes, $n$ and not $d$) from $\boldsymbol{D}$ such that:

- $\sum_{i=1}^n \boldsymbol{x}_i = \mathbf{0}$;
- $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \geq D_{i,j}$ for all $i \neq j$; and
- $\max\left\{\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2}{D_{i,j}} : 1 \leq i < j \leq n\right\}$ is as small as possible.

*Hint*: use semidefinite programming.

*Solution.* $\square$

**Problem 4** (25 points)**.** Exercise 3.25 from BHK.

*Solution.* ☐