Homework 12
HDP KNU+ Fall 2022

Hints are in the back of this homework set.

The *Gaussian mixture model* is a popular mathematical model of high-dimensional structured data. It consists of two or more clusters formed by points in $\mathbb{R}^d$ sampled from Gaussian distributions with different means.

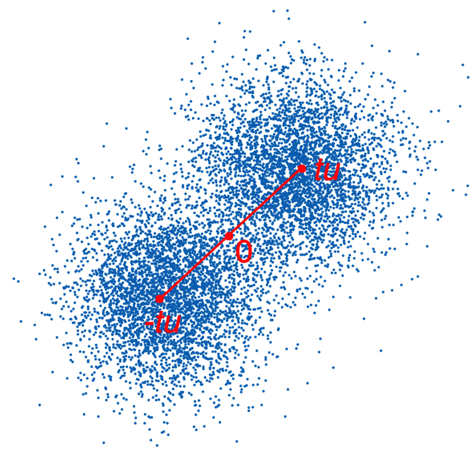

Figure 1. Gaussian mixture model

The Gaussian mixture model with two clusters in $\mathbb{R}^d$ is illustrated in Figure 1. To generate it, we can flip an independent fair coin, and depending on the outcome sample a point either from $N(tu, I_n)$ or $N(-tu, I_d)$. Here $u \in \mathbb{R}^d$ is a unit vector that determines the direction in which the centers of the clusters are separated, and $t > 0$ determines the strength of separation. Repeating the process n times, we get the data points $X_1, \ldots, X_n \in \mathbb{R}^d$.

We can formalize the model by considering an independent random variable θ that takes values -1 and 1 with probability $1/2$ each. Let $Z \sim N(0, I_d)$, and define

$$X = Z + \theta t u.$$

Let $X_i = Z_i + \theta_i t u$, $i = 1, \ldots, n$, be independent random vectors drawn from the same distribution as X. Assume that

$$\|u\|_2 = 1 \quad \text{and} \quad t \geq 0.1.$$

Let us show how to learn the model, namely estimate the direction of separation u, from a sample of size $n = O(d)$.

Problem 1 (Learning the gaussian mixture model)

(a) Let Y and Z are independent mean zero random vectors in $\mathbb{R}^d$. Prove that

$$\Sigma(Y + Z) = \Sigma(Y) + \Sigma(Z),$$

where $\Sigma(X) = \mathbb{E}\, XX^\mathsf{T}$ denotes the covariance matrix of a mean zero random vector X.

(b) Compute the covariance matrix of the Gaussian mixture model, showing that

$$\Sigma(X) = I_d + t^2 uu^\mathsf{T}.$$

(c) Check that the top eigenvalues and eigenvector of $\Sigma = \Sigma(X)$ are

$$\lambda_1(\Sigma) = 1 + t^2, \quad \lambda_2(\Sigma) = 1, \quad v_1(\Sigma) = u.$$

(d) Let $v = v_1(\Sigma_n)$ and $\lambda = \lambda_1(\Sigma_n)$, where Σ_n is the sample covariance matrix of the data. Show that if $n > Cd$ (where C is a sufficiently large absolute constant), then with probability at least 0.9 we have

$$\|u - v\|_2 \leq 0.1.$$

Thus, we can estimate the direction of separation u from the data $X_1, \ldots, X_n$.

The general covariance estimation result (Lecture 33, November 18) guarantees that the covariance of a d-dimensional distribution can be accurately estimated from a sample of size $n = O(d \log d)$. In this problem, we show that the logarithmic oversampling is in general necessary.

PROBLEM 2 (LOGARITHMIC OVERSAMPLING)

(a) Check that all entries of a matrix are bounded by the operator norm of the matrix, i.e.

$$\max_{i,j} |A_{ij}| \leq \|A\|.$$

(b) Let $e_1, \ldots, e_d$ be the canonical basis of $\mathbb{R}^d$. Consider a random vector X that is uniformly distributed in the set $\{\sqrt{d}\, e_1, \ldots, \sqrt{d}\, e_d\}$, i.e. X takes each of the n values $\sqrt{d}\, e_k$ with probability $1/d$. Check that

$$\Sigma = \mathbb{E}\, XX^\mathsf{T} = I_d.$$

(c) Argue that, unless $n > cd \log d$, the sample covariance matrix Σ_n has at least one zero diagonal entry with probability at least 0.9.

In solving this part, you can assume without proof the following classical result in probability about the *coupon collector's problem.* Suppose each day we receive a coupon in mail. There are d types of coupons, and each day's coupon is one of these types independently with equal probability. Then, with probability at least 0.9, we need at least $cd \log d$ days to obtain a full collection of all d types of coupons.

(d) Conclude that if $n < cd \log d$, we have $\|\Sigma_n - \Sigma\| \geq \|\Sigma\|$ with probability at least 0.9, so the covariance estimation fails.

The general covariance estimation result (Lecture 33, November 18) has a strange requirement that $\|X\|_2^2 \lesssim \mathbb{E}\|X\|_2^2$ with probability 1. In the following problem, we argue that this requirement can not be removed. even in dimension $d = 1$.

Problem 3 (Boundedness requirement)

Fix any integer N. Construct a random variable X whose variance satisfies $\Sigma := \mathbb{E}\, X^2 = 1$ but, if $n < N$, the sample covariance satisfies $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i^2 = 0$ with probability 0.9. (Here X_i denote independent random variables sampled from the same distribution as X.)

Some real-world problems require us to operate with huge matrices (for example, matrices of interactions between hundreds of thousands of individuals). Some matrices are so huge that computing their eigenvectors and eigenvalues is impossible. In such cases, one can reduce the size of a matrix by *subsampling*: including only a random sample of rows or columns.

In the following problem, we consider a very tall $N \times d$ matrix A, with $N \gg n$. We show how to approximately compute the singular values of A from a smaller matrix B obtained by subsampling $n = O(d \log d)$ rows of A.

Problem 4 (Subsampling a matrix)

Let A be a $N \times d$ matrix whose rows A_i^T have the same length, i.e. $\|A_i\|_2 = \|A_j\|_2$ for all $i = 1, \ldots, N$. Consider the $n \times d$ matrix R obtained by including n randomly chosen rows of A, where each time a row is chosen independently, with replacement, and with the same probability $1/N$. Let $B = \sqrt{N/n}\, R$. Show that, as long as $n \geq Cd \log d$ with a sufficiently large absolute constant C, the singular values satisfy

$$\max_{i=1,\ldots,n} \left| \sigma_i(A)^2 - \sigma_i(B)^2 \right| \leq 0.1\, \sigma_1(A)^2$$

with probability at least 0.9.

TURN OVER FOR HINTS

Hints

Hints for Problem 1. (b) follows from (a) with $Y = \theta t u$.

(d) Apply the covariance estimation result from Homework 9, Problem 3(b), followed by Davis-Kahan inequality (Homework 10, Problem 1(c)).

Hints for Problem 2. (d) The matrix Σ_n has a zero diagonal entry by (c), but the corresponding entry of Σ is 1 by (b). The result now follows from (a).

Hints for Problem 3. Consider $X \sim \mathrm{Ber}(p)$. If the parameter p of the Bernoulli distribution is very small, the entire sample is zero with high probability.

Hints for Problem 4. Consider a random vector $X \in \mathbb{R}^d$ that is uniformly distributed on the set of rows of A, i.e. X takes values A_k^T with probability $1/N$ for each $k = 1, \dots, n$. Check that $\Sigma = \frac{1}{N} A^\mathsf{T} A$ and $\Sigma_n = \frac{1}{n} R R^\mathsf{T}$. Apply the general covariance estimation result (Lecture 33, November 18) to get $\|A^\mathsf{T} A - B^\mathsf{T} B\| \le 0.1 \, \|A^\mathsf{T} A\|$. Finally, use Weyl's inequality (Lecture 24, October 26) and recall how singular values of A and B are related to eigenvalues of $A^\mathsf{T} A$ and $B^\mathsf{T} B$.