Lecture 10

Review * Given a matrix $A$ with each row denoting a $d$-dimensional data vector, let $A_k$ denote the matrix obtained by projecting each row of $A$ on the space spanned by right-singular vectors with $k$ largest singular values. Then, for any $k$-rank matrix $B$,

$$\|A - A_k\|_F \leq \|A - B\|_F,$$

and

$$\|A - A_k\|_2 \leq \|A - B\|_2.$$

Agenda * (Contd.) Learning Gaussian mixtures

- Vempala-Wang projection for learning the span of the means $\{\mu_1, ..., \mu_k\}$.

1. SVD for estimating the span of means

$X_1, ..., X_n$ are iid from $\sum_{j=1}^n \omega_j N(\mu_j, \sigma^2 I_d \otimes I_d)$

Notation: $A$ be the $n \times d$ matrix with the $j$th row $x_j$

- Given a space $U$, denote the projection of $x$ on $U$ by $\text{proj}_U x$ and the matrix obtained by projecting each row of $A$ on $U$ by $\text{proj}_U A$.

$$\|\text{proj}_U A\|_F = \sum_{i=1}^n \|\text{proj}_U A_i\|_2.$$ 

What will we show? (1) On average, the best $k$-dim space approximating $A$ is $U = \text{span} \{\mu_1, ..., \mu_k\}$.
(2) With large prob., most of the energy of $U$ is along $V$, the space spanned by the top $k$ right singular values of $A$.

**Theorem**

Let $U = \text{span}\{u_1, \ldots, u_k\}$.

Let $V$ be a linear space with $\dim(V) = \dim(U)$.

Then,

$$\mathbb{E}\left[\|\text{proj}_U A\|_F^2\right] \geq \mathbb{E}\left[\|\text{proj}_V A\|_F^2\right].$$

**Proof.** (a) Let $X = (X_1, \ldots, X_n)$ consist of uncorrelated entries.

Let $\mu = \mathbb{E}[X]$ and $\text{Var}(X_i) = \sigma^2$. Then,

$$\mathbb{E}\left[(X_i u_j)^2\right] = \sum_{i \neq j} \mathbb{E}\left[X_i X_j u_i u_j\right] + \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] u_i^2$$

$$= \sum_{i \neq j} \mu_i \mu_j u_i u_j + \sum_{i=1}^{n} (\mu_i^2 + \sigma^2) u_i^2$$

$$= (\mu, u)^2 + \sigma^2 \|u\|_2^2$$

(b) For any $r$ dimensional space $V$,

$$\mathbb{E}\left[\|\text{proj}_V X\|_F^2\right] = \|\text{proj}_V \mathbb{E}(X)\|_F^2 + \sigma^2 \|u\|_2^2$$

Indeed, let $V_1, \ldots, V_n$ be an o.n. for $V$. Then,

$$\text{proj}_V X = \sum_{i=1}^{n} (X_i v_i) v_i$$

and

$$\|\text{proj}_V X\|_F^2 = \sum_{i=1}^{n} (X_i v_i)^2.$$ 

Thus,

$$\mathbb{E}\left[\|\text{proj}_V X\|_F^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(X_i v_i)^2\right].$$
\[  = \sum_{i=1}^{\pi} (E[X_i, \mathcal{V}_i])^2 + n \sigma^2 \]
\[  = \| \text{proj}_U E[X] \|_2^2 + n \sigma^2. \]

(c) Let \( A \) be the data matrix as before, generated from a mixture \( \sum_{j=1}^{k} \omega_j P_j \), where \( (\mu_j, \sigma_j, I) \) denote the mean and covariance matrix for \( X_i \).

Then,
\[ E[\| \text{proj}_U A \|_F^2] = n \sum_{j=1}^{k} \omega_j \left( \| \text{proj}_U \mu_j \|_2^2 + n \sigma_j^2 \right) \]

The relation above can be seen as follows:

Let \( N_j \) denote the number of samples from \( P_j \).

Thus,
\[ E \left[ \| \text{proj}_U A \|_2^2 \right] = E \left[ \sum_{j=1}^{k} N_j \| \text{proj}_U Y_j \|_2^2 \right] \]
\[ \text{where } Y_j \sim P_j \]
\[ = E \left[ \sum_{j=1}^{k} N_j \left( \| \text{proj}_U \mu_j \|_2^2 + n \sigma_j^2 \right) \right] \]
\[ = \sum_{j=1}^{k} n \omega_j \left( \| \text{proj}_U \mu_j \|_2^2 + n \sigma_j^2 \right) \]

(d) Finally, we prove the theorem.

\[ E \| \text{proj}_U A \|_2^2 - E \| \text{proj}_U A \|_F^2 \]
\[ = n \sum_{j=1}^{k} \omega_j \left( \| \text{proj}_U \mu_j \|_2^2 - \| \text{proj}_U \mu_j \|_F^2 \right) \]
\[ = n \sum_{j=1}^{k} \omega_j \left( \| \mu_j \|_2^2 - \| \text{proj}_U \mu_j \|_2^2 \right) \geq 0. \]
Remark. Note that while span \{\mu_1, \ldots, \mu_k\} captures the energy along the means, noise energy is spread evenly in all directions and any extra dimension used in \( V \) will capture it better.

**Theorem.** Let \( V \) denote the \( k \)-dimensional space spanned by the top \( k \) right-singular vectors of \( A \).

Then, if \( n = \mathcal{O}\left( \frac{d}{\sigma^2 \omega_{\text{min}}} \right) \), with large prob.

\[
\sum_{i=1}^{k} \omega_i \left( \| \mu_i \|_2^2 - \| \text{proj}_V \mu_i \|_2^2 \right) \leq \delta \left( d-k \right) \sum_{j=k+1}^{d} \omega_j \sigma_j^2.
\]

**Proof.** Involved. We need the following concentration result:

For a \( k \)-dimensional space \( V \) and \( X \sim N(\mu, \sigma^2 I) \),

\[
P \left( \| \text{proj}_V X \|_2 > (1+\epsilon) \mathbb{E} \left[ \| \text{proj}_V X \|_2^2 \right] \right) < e^{-\epsilon^2 k/8}
\]

\[
P \left( \| \text{proj}_V X \|_2 < (1-\epsilon) \mathbb{E} \left[ \| \text{proj}_V X \|_2^2 \right] \right) < e^{-\epsilon^2 k/8}
\]

\[
\rightarrow \text{Let's see an easy version:}
\]

\[
P \left( (X, \sigma)^2 > (1+\epsilon)(\mu_0^2 + \sigma^2 \| \Sigma \|_2^2) \right)
\]

Assume \( \Sigma \) is unit norm. Then, \( (X, \Sigma) \) is a Gaussian with mean \( (\mu, \Sigma) \) and variance \( \sigma^2 \). Thus, the required prob. is simply \( P \left( Z^2 > (1+\epsilon)(\theta^2 + \sigma^2) \right) \) for \( Z \sim N(\theta, \sigma^2) \).

This concentration bound for Chi-square distribution is known. \( \blacksquare \)
Learning Gaussian mixtures

→ Distance based clustering

* We use first $\tilde{O}(d^4)$ samples and find the
  space $V_k$, the space spanned by the top $k$ singular
  vectors of $A$.

* Now, take another set of $n$ samples and form a new
  $A$. Project each row of $\tilde{A}$ on $V_k$.

Note that the projected samples are $k$ dimensional
and have means $\mu_i', \mu_j'$ satisfying

$$\|\mu_i' - \mu_j'\|_2^2 \geq \|\mu_i - \mu_j\|_2^2 - \delta d \sigma^2$$

(ending the second

If we choose $\delta = \frac{1}{d}$ (use $\tilde{O}(d^4)$ samples),
we have $\|\mu_i' - \mu_j'\|_2^2 \geq \|\mu_i - \mu_j\|_2^2 - \sigma^2$,
which will allow us to use distance based clustering
to distinguish clusters if $\|\mu_i - \mu_j\|_2 = \Omega(\sqrt{k} \sigma^2)$.

→ Sphere for learning distributions

Quantize the coefficients $a_1, ..., a_k$ of the parameterized
space $V_k = \{ \sum_{i=1}^k \alpha_i v_i \}, (\alpha_1, ..., \alpha_k) \in \mathbb{R}^k$.

To limit our guesslist, we need to start with a bound
on $\max_i \|\mu_i - \mu_j\|_2$.  