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,

$$\|A - A_k\|_F \leq \|A - B\|_F, \text{ and}$$

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

Agenda * (Contd.) Learning Gaussian mixtures

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

$\boxed{A}$ SVD for estimating the span of means

$X_1, \dots, X_n$ are iid from $\sum_{j=1}^{n} w_j \, \mathcal{N}(\mu_j, \sigma^2 I_{d\times d})$

Notations → 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 $\text{proj}_U A$.

→ $\|\text{proj}_U A\|_F^2 = \sum_{i=1}^{n} \|\text{proj}_U A_i\|_2^2$

What will we show? (1) On average, the best k-dim space approximating A is $U = \text{span}\{\mu_1, \dots, \mu_k\}$

(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.

2

Theorem Let $U = \text{span}\{\mu_1, \dots, \mu_k\}$.

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

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 $X = (X_1, \dots, X_n)$ consist of uncorrelated entries.

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

$$\mathbb{E}\left[(X.v)^2\right] = \sum_{i,j} \mathbb{E}\left[X_i X_j v_i v_j\right]$$
$$= \sum_{i \neq j} \mathbb{E}[X_i]\mathbb{E}[X_j] v_i v_j + \sum_{i=1}^{n} \mathbb{E}[X_i^2] v_i^2$$
$$= \sum_{i \neq j} \mu_i \mu_j v_i v_j + \sum_{i=1}^{n} (\mu_i^2 + \sigma^2) v_i^2$$
$$= (\mu . v)^2 + \sigma^2 \|v\|_2^2$$

(b) For any r dimensional space V,
$$\mathbb{E}\left[\|\text{proj}_V X\|_2^2\right] = \|\text{proj}_V \mathbb{E}[X]\|_2^2 + \sigma^2 r$$

Indeed, let $v_1, \dots, v_r$ be an o.n. for V. Then,

$$\text{proj}_V X = \sum_{i=1}^{r} (X.v_i) v_i \quad \text{and}$$
$$\|\text{proj}_V X\|_2^2 = \sum_{i=1}^{r} (X.v_i)^2.$$

Thus, $$\mathbb{E}\left[\|\text{proj}_V X\|_2^2\right] = \sum_{i=1}^{r} \mathbb{E}\left[(X.v_i)^2\right]$$

③

$$= \sum_{i=1}^{r} (\mathbb{E}[X].v_i)^2 + r\sigma^2$$

$$= \|\mathrm{proj}_V \mathbb{E}[X]\|_2^2 + r\sigma^2.$$

(c) Let A be the data matrix as before, generated from a mixture $\sum_{j=1}^{k} w_j P_j$, where $(\mu_j, \sigma_j^2 I)$ denote the mean and covariance matrix for X_i.

Then,

$$\mathbb{E}\left[\|\mathrm{proj}_V A\|_F^2\right] = n\sum_{j=1}^{k} w_j \left(\|\mathrm{proj}_V \mu_j\|_2^2 + r\sigma_j^2\right)$$

The relation above can be seen as follows:

Let N_j denote the number of samples from P_j.

Thus, $\mathbb{E}\left[\|\mathrm{proj}_V A\|_2^2\right] = \mathbb{E}\left[\sum_{j=1}^{k} N_j \ \mathbb{E}\|\mathrm{proj}_V Y_j\|^2\right]$

where $Y_j \sim P_j$

$$= \mathbb{E}\left[\sum_{j=1}^{k} N_j\left(\|\mathrm{proj}_V \mu_j\|_2^2 + r\sigma_j^2\right)\right]$$

$$= \sum_{j=1}^{k} n w_j\left(\|\mathrm{proj}_V \mu_j\|_2^2 + r\sigma_j^2\right)$$

(d) Finally, we prove the theorem.

$\mathbb{E}\|\mathrm{proj}_U A\|_2^2 - \mathbb{E}\|\mathrm{proj}_V A\|_2^2$

$$= n\sum_{j=1}^{k} w_j \left(\|\mathrm{proj}_U \mu_j\|_2^2 - \|\mathrm{proj}_V \mu_j\|_2^2\right)$$

$$= n\sum_{j=1}^{k} w_j \left(\|\mu_j\|_2^2 - \|\mathrm{proj}_V \mu_j\|_2^2\right) \geq 0.$$

∎

<u>Remark</u>. Note that while span$\{\mu_1, \dots, \mu_k\}$ 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.

<u>Theorem</u> Let V denote the k-dimensional space spanned by the top k right-singular vectors of A.
Then, if $n = \tilde{O}\left(\frac{d}{\delta^2 \omega_{min}}\right)$, with large prob.

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

<u>Proof</u>. Involved. We need the following concentration result:
For a k-dimensional space V and $X \sim \mathcal{N}(\mu, \sigma^2 I)$,

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

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

→ Let's see an easy version:

$$\mathbb{P}\left((X.v)^2 > (1+\epsilon)\left((\mu.v)^2 + \sigma^2 \|v\|^2\right) \right)$$

Assume v is unit norm. Then, $(X.v)$ is a Gaussian with mean $(\mu.v)$ and variance σ^2. Thus, the required prob. is simply $\mathbb{P}\left(Z^2 > (1+\epsilon)(\theta^2 + \sigma^2)\right)$ for $Z \sim \mathcal{N}(\theta, \sigma^2)$.
This concentration bound for Chi-square distribution is known. ∎

B Learning Gaussian mixtures

→ Distance based clustering

* We use first $\tilde{O}\left(\frac{d}{\delta^2}\right)$ 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 $\tilde{A}$. Project each row of $\tilde{A}$ on V_k.

  Note that the projected samples are k dimensional and have means μ_i', μ_j' satisfying

  $\|\mu_i' - \mu_j'\|_2^2 \geq \|\mu_i - \mu_j\|_2^2 - \delta d \sigma^2$ (using the second theorem)

If we choose $\delta = \frac{1}{d}$ (use $\tilde{O}(d^3)$ 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 $\|\mu_i - \mu_j\|_2^2 = \Omega(\sqrt{k}\,\sigma^2)$.

→ Scheffé for learning distributions

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

To limit our guesslist, we need to start with a bound on $\max_{i,j} \|\mu_i - \mu_j\|_2$.