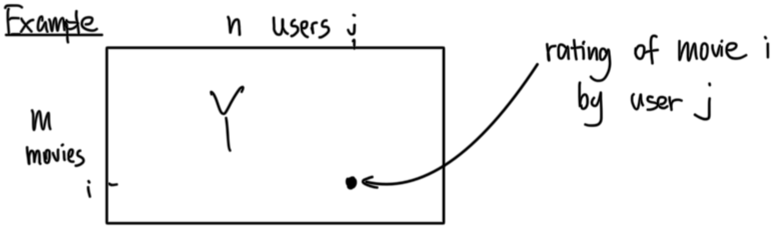
Lecture 4

Matrix factorization models (single-spike model)

Example



Statistical model

$$\underline{Y}_{ij} = \underbrace{a_i^0}_{\text{quality of movie } i} \cdot \underbrace{b_j^0}_{\text{user rating strictness}} + \underbrace{\underline{W}_{ij}}_{\text{external factors noise}} \quad , \quad \underline{W}_{ij} \sim N(0,1)$$

matrix form $\underline{Y} = X^0 + \underline{W}$, for unknown rank-1 matrix $X^0 = a^0 (b^0)^T$.

goal: estimate X^0 given $\underline{Y}$

error measure: $\frac{1}{nm} \| \hat{X} - X^0 \|_F^2 = \frac{1}{nm} \sum_{i,j} (\hat{X}_{ij} - x_{ij}^0)^2$

thought experiment:

Suppose we know b^0, least squares estimator is

$$\hat{\underline{a}} = \operatorname{argmin}_a \| a \cdot (b^0)^T - \underline{Y} \|_F^2$$

* Sounds like cheating?!

error bound:

$$\mathbb{E}\left[\frac{1}{n \cdot m} \| a^0 (b^0)^T - \hat{a} (b^0)^T \|_F^2 \right] \leq \frac{m}{n \cdot m} = \frac{1}{n}$$

minimax optimality of LS $\longrightarrow$ lower bound, also for original prob.

by symmetry, error lower bound is $\max \{ \frac{1}{m}, \frac{1}{n} \}$.

Maximum likelihood estimator

Best rank-1 approximation:

$$\hat{\underline{X}} = \operatorname{argmin} \{ \underbrace{\| X - \underline{Y} \|_F^2}_{\text{convex in } X} \mid \underbrace{\operatorname{rank}(X) \leq 1}_{\text{non-convex}} \}$$

* There exists poly time alg. to compute $\hat{\underline{X}}$.

- $\min \| ab^T - \underline{Y} \|_F^2$
- $a \in \mathbb{R}^m$
- $b \in \mathbb{R}^n$

$$\boxed{\text{Theorem} \quad \mathbb{E}\left[\frac{1}{nm} \| \hat{\underline{X}} - X^0 \|_F^2 \right] \leq O\left(\frac{1}{n} + \frac{1}{m} \right)}$$

Interlude: Singular vector decomposition (SVD)

$$Y = \sum_{i=1}^{r} \sigma_i \cdot u_i \cdot v_i \quad , \quad \text{rank}(Y) = r$$

- singular values: $\sigma_1 \geq \cdots \geq \sigma_r > 0$
- left singular value vectors: $u_1, \cdots, u_r$ orthonormal
- right singular value vectors: $v_1, \cdots, v_r$ orthonormal

<u>fact</u>: SVD always exists and poly-time computable

<u>fact</u>: $\sigma_1 \cdot u_1 \cdot v_1^T$ is the best rank-1 approx. to Y * The first term

<u>matrix norms</u>:

- Spectral norm: $\|X\| = \sigma_1$
- Nuclear norm: $\|X\|_* = \sigma_1 + \cdots + \sigma_r$
- Frobenius norm: $\|X\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$

<u>inner product</u>

$$\langle X, Y \rangle = \sum X_{ij} Y_{ij} = \text{Tr}(XY^T)$$

$$\langle Y, Y \rangle = \|Y\|_F^2$$

<u>inequalities</u>

- $\|Y\| \leq \|Y\|_F \leq \|Y\|_*$
- $\|Y\|_* \leq \sqrt{r} \cdot \|Y\|_F$
- $\langle X, Y \rangle \leq \|X\| \cdot \|Y\|_*$

<u>theorem</u> SVD

$W_{ij} \sim N(0,1)$, $\mathbb{E}[\|W\|^2] \leq O(n+m)$

<u>Proof</u>: SVD

$\underline{Y} = X^0 + \underline{W}$, $\text{rank}(X^0) = 1$, $\underline{W}_{ij} \overset{iid}{\sim} N(0,1)$

$\hat{X} = \text{argmin}\{\|X - \underline{Y}\|_F^2 \mid \text{rank}(X) = 1\}$

$\underline{U} = \hat{\underline{X}} - X^0 \longrightarrow \text{rank}(\underline{U}) \leq \text{rank}(\hat{\underline{X}}) + \text{rank}(X^0) \leq 2$

<u>claim 1</u>: $\|\underline{U}\|_F^2 \leq 2\langle \underline{U}, \underline{W} \rangle$ * same as for BSS, Huber loss, ...

<u>Hölder</u>: $\langle \underline{U}, \underline{W} \rangle \leq \|\underline{U}\|_* \cdot \|\underline{W}\|$

<u>claim 2</u>: $\|\underline{U}\|_* \leq \sqrt{2} \cdot \|\underline{W}\|_F$

* $W_{ij} \sim N(0,1)$

Gordon's theorem for Gaussian matrix

$\mathbb{E}[\|W\|] \lesssim \sqrt{m} + \sqrt{n}$, the expected value of the largest singular value of a $m \times n$ standard Gaussian is bounded by the sum of square roots of its dimensions.

<u>together</u>:

$$\|\underline{U}\|_F^2 \overset{\text{claim 1}}{\leq} 2\langle \underline{U}, \underline{W} \rangle \overset{\text{Hölder}}{\leq} 2\|\underline{U}\|_* \cdot \|\underline{W}\| \overset{\text{claim 2}}{\leq} 2\sqrt{2}\,\|\underline{U}\|_F \cdot \|\underline{W}\|$$

$\rightarrow \|\underline{U}\|_F^2 \leq 8\|\underline{W}\|^2$

$\rightarrow \mathbb{E}[\|\underline{U}\|_F^2] \leq 8\,\mathbb{E}[\|\underline{W}\|^2] \leq O(n+m)$

<u>Multi-spike matrix model</u>

$Y = X^0 + W$, $W_{ij} \sim N(0,1)$, unknown X^0, $\text{rank}(X^0) \leq r$.

Motivation

$Y_{ij} = \left(\sum_k a^\circ_{ik} \cdot b^\circ_{jk}\right) + W_{ij} \longrightarrow X^\circ = A^\circ \cdot (B^\circ)^T$

- r: archetypes of users
- a°_{ik}: quality of movie i for user archetype k
- user j: combination of all r archetypes with weights $b^\circ_{j1}, \cdots, b^\circ_{jr}$

MLE

$$\hat{X} = \arg\min \{ \|X - Y\|_F^2 \mid \mathrm{rank}(X) \le r \}$$
$$= \sigma_1 \cdot u_1 \cdot v_1^T + \cdots + \sigma_r \cdot u_r \cdot v_r^T$$ * First r terms of SVD of Y

Theorem

$$\mathbb{E}\left[\frac{1}{nm}\|\hat{X} - X^\circ\|_F^2\right] \le r \cdot O\left(\frac{1}{n} + \frac{1}{m}\right)$$

Single spike vs. multiple spike

Single spike: $Y = a^*(b^*)^T + W$, $a^* \in \mathbb{R}^m$, $b^* \in \mathbb{R}^n$

(rank-1) $Y_{ij} = a^*_i (b^*_j)^T + W_{ij}$

$$\begin{bmatrix} Y_{11} & \cdots & Y_{1n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \cdots & Y_{mn} \end{bmatrix} = \begin{bmatrix} a^*_1 \\ \vdots \\ a^*_m \end{bmatrix} \begin{bmatrix} b^*_1 & \cdots & b^*_n \end{bmatrix} + W$$

Multiple spike: $Y = A^*(B^*)^T + W$

(rank-r) $A^* \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{n\times r}$

$$Y_{ij} = \sum_{k=1}^r A^*_{ik} B^*_{jk} + W_{ij} = \langle A^*_{i\cdot}, B^*_j \rangle + W_{ij}$$

$$\begin{bmatrix} Y_{11} & \cdots & Y_{1n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \cdots & Y_{mn} \end{bmatrix} = \begin{bmatrix} A^*_{11} & \cdots & A^*_{1r} \\ \vdots & \ddots & \vdots \\ A^*_{m1} & \cdots & A^*_{mr} \end{bmatrix} \begin{bmatrix} B^*_{11} & \cdots & B^*_{n1} \\ \vdots & & \vdots \\ B^*_{1r} & \cdots & B^*_{nr} \end{bmatrix} + W$$

Matrix completion

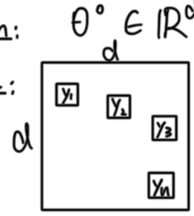
Allow missing data

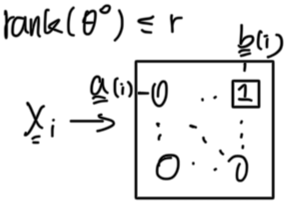
New notation!

parameters: rank $r \ge 2$, dimension $d \ge 2$, sample size n

unknown: $\theta^\circ \in \mathbb{R}^{d\times d}$, $\mathrm{rank}(\theta^\circ) \le r$

observe:





notation: $(X_1, y_1), \cdots, (X_n, y_n)$

- $y_i = \langle X_i, \theta^\circ \rangle + w$
- $X_i = e_{a(i)} \cdot e^T_{b(i)}$
  - $a(1), \cdots, a(n), b(1), \cdots, b(n) \overset{iid}{\sim} \mathrm{Unif}(\{1, \cdots, d\})$
  - $w_1, \cdots, w_n \sim N(0,1)$

MLE $\hat{X} = \arg\min\left\{ \sum_{i=1}^n (\langle X_i, \theta \rangle - y_i)^2 \mid \mathrm{rank}(\theta) \le r \right\}$