Non-negative matrix factorization

Daniel Hsu

COMS 4772

Singular value decomposition

- $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$
  - $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$
  - $\boldsymbol{S} \succ 0$ diagonal
  - Truncations at rank k are optimal for spectral/Frobenius error
- What if we want to add constraints to factors?

Non-negative matrix factorization (NMF)

- **Given**: $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ non-negative
  - Columns are, e.g., word frequencies of documents, pixel intensities of images.
- **Goal**: factor $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{Y}$ where $\boldsymbol{V} \in \mathbb{R}^{m \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{r \times n}$ have only non-negative entries
  - NP-hard to decide if this is possible (Vavasis, 2007)

Heuristic (Lee & Seung, 1999)

- Write approximation objective $f(\boldsymbol{V}, \boldsymbol{Y}) := \|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{Y}\|_F^2$ as

$$\begin{aligned} f(\boldsymbol{V}, \boldsymbol{Y}) &= \sum_{i,j} X_{i,j}^2 - 2X_{i,j}(\boldsymbol{U}\boldsymbol{V})_{i,j} + (\boldsymbol{V}\boldsymbol{Y})_{i,j}^2 \\ &= \underbrace{\|\boldsymbol{X}\|_F^2 + \|\boldsymbol{V}\boldsymbol{Y}\|_F^2}_{\geq 0} - \underbrace{2\,\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{V}\boldsymbol{Y})}_{\geq 0} \\ &= f_+(\boldsymbol{V}, \boldsymbol{Y}) - f_-(\boldsymbol{V}, \boldsymbol{Y}) \,. \end{aligned}$$

- **Multiplicative updates** (preserves non-negativity):

$$V_{i,k} \leftarrow V_{i,k} \cdot \frac{\frac{\partial}{\partial V_{i,k}} f_-(\boldsymbol{V}, \boldsymbol{Y})}{\frac{\partial}{\partial V_{i,k}} f_+(\boldsymbol{V}, \boldsymbol{Y})} \,, \qquad Y_{k,j} \leftarrow Y_{k,j} \cdot \frac{\frac{\partial}{\partial Y_{k,j}} f_-(\boldsymbol{V}, \boldsymbol{Y})}{\frac{\partial}{\partial Y_{k,j}} f_+(\boldsymbol{V}, \boldsymbol{Y})}$$

- Update factor ≥ 1 iff $f'(\boldsymbol{V}, \boldsymbol{Y}) \leq 0$.
- **Fixed points**: $\boldsymbol{V} = \boldsymbol{0}$, $\boldsymbol{Y} = \boldsymbol{0}$, or stationary point of f.

Example

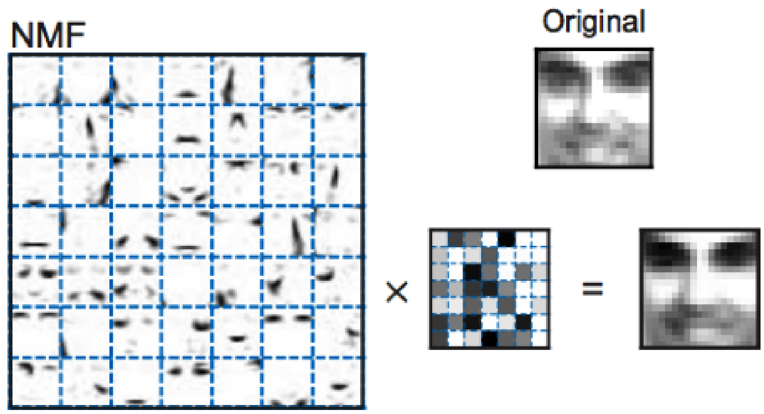

Figure 1: NMF for face images

Recovery problem

- Suppose $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{Y}$ for some non-negative $\boldsymbol{V}$ and $\boldsymbol{Y}$ of rank r.
  - Assume (WLOG) rows of $\boldsymbol{X}$, $\boldsymbol{V}$, and $\boldsymbol{Y}$ sum to 1.
  - Each row of $\boldsymbol{X}$ is a convex combination of rows of $\boldsymbol{Y}$.
- **Given**: $\boldsymbol{X}$.
- **Goal**: recover factors $\boldsymbol{V}$ and $\boldsymbol{Y}$.
- **Separability assumption**: $\boldsymbol{V}$ has positive definite diagonal submatrix.
  - Ensures uniqueness (Donoho & Stodden, 2003; Arora, Ge, Kannan, & Moitra, 2012)
  - Each row of $\boldsymbol{Y}$ appears as a row of $\boldsymbol{X}$ (possibly scaled).
  - (Scaling factor is 1 under assumption that rows of $\boldsymbol{V}$ sum to 1.)

Recovery algorithm (Arora, Ge, Kannan, & Moitra, 2012)

- **Main idea**: identify the rows of $\boldsymbol{X}$ that are exactly rows of $\boldsymbol{Y}$.
- For each $i = 1, 2, \ldots, m$:
  - If i-th row of $\boldsymbol{X}$ is in convex hull of all other rows of $\boldsymbol{X}$, then delete the i-th row of $\boldsymbol{X}$
- What remains is exactly r rows of $\boldsymbol{X}$, each being a row of $\boldsymbol{Y}$.

Application: topic models

- $X_{w,d}$ = number of times word w appears in document d
- $V_{w,t} = \Pr(\text{word } w \mid \text{topic } t)$
- $Y_{t,d} \propto \Pr(\text{topic } t \mid \text{document } d)$
- $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{V}\boldsymbol{Y}$
- **Separability assumption**: for every topic t, there is a word w_t that has non-zero probability in $\boldsymbol{V}$ only under topic t.
  - E.g., word “backprop” only appears in documents about topic “machine learning”
- **Goal**: estimate $\boldsymbol{V}$ from documents
  - When model is well-specified,

$$\boldsymbol{X} = \boldsymbol{V}\boldsymbol{Y} + \text{zero-mean noise}.$$

Using co-occurrences (Arora, Ge, & Moitra, 2012)

- Assume each document has two tokens (i.e., length ≥ 2)
- **Bag-of-words assumption with $(\boldsymbol{V}, \boldsymbol{Y})$ model**: for document d,
  - First token is word w with probability $\sum_t V_{w,t} Y_{t,d}$
  - Second token is word w with probability $\sum_t V_{w,t} Y_{t,d}$ (independent of first token)
- **Co-occurrence matrix**: $M_{w,w'}$ = number of documents where first token is w and second token is w'.

$$\mathbb{E}(\boldsymbol{M}) \; = \; \boldsymbol{V}\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{V}^\top \, .$$

  - Separability of $\boldsymbol{V}$ can be used with $\mathbb{E}(\boldsymbol{M})$.
  - If documents are independent, then $\boldsymbol{M}$ is sum of independent random matrices; can exploit matrix concentration to bound $\|\boldsymbol{M} - \mathbb{E}(\boldsymbol{M})\|_2$.