Decompositions of moment tensors

Daniel Hsu

COMS 4772

Tensor decompositions

High-dimensional support recovery from moments

- Random vector $\boldsymbol{X}$, supported on k distinct points $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k \in \mathbb{R}^d$,

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{z}_t) \;=\; w_t > 0\,.$$

- Can we learn the parameters $\{(w_t, \boldsymbol{z}_t)\}_{t=1}^k$ from moments?
- **Moments**:

$$\begin{aligned}
\mathbb{E}(\boldsymbol{X}) &= \sum_{t=1}^k w_t \cdot \boldsymbol{z}_t\,, \\
\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) &= \sum_{t=1}^k w_t \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t\,, \\
\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X} \otimes \boldsymbol{X}) &= \sum_{t=1}^k w_t \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t \otimes \boldsymbol{z}_t\,.
\end{aligned}$$

Insufficiency of low-order moments

- The following looks like eigenvalue decomposition:

$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \sum_{t=1}^k w_t \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t\,,$$

  but $\{\boldsymbol{z}_t\}_{t=1}^k$ need not be orthogonal.
- Possible to have

$$\sum_{t=1}^k w_t \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t \;=\; \sum_{t=1}^k \tilde{w}_t \cdot \tilde{\boldsymbol{z}}_t \otimes \tilde{\boldsymbol{z}}_t$$

  for different parameters $\{(w_t, \boldsymbol{z}_t)\}_{t=1}^k$ and $\{(\tilde{w}_t, \tilde{\boldsymbol{z}}_t)\}_{t=1}^k$.
- **Note**: additional constraints *could* make decomposition unique (e.g., separability in NMF).

Jennrich's algorithm (1970)

- Define

$$\boldsymbol{S} \;:=\; \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \sum_{t=1}^{k} w_t \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t \,,$$

$$\boldsymbol{T} \;:=\; \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \sum_{t=1}^{k} w_t \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t \otimes \boldsymbol{z}_t \,.$$

- **Key assumption**: $\boldsymbol{Z} := [\boldsymbol{z}_1 | \boldsymbol{z}_2 | \cdots | \boldsymbol{z}_k]$ has rank k.
- **Main idea**: combine $\boldsymbol{S}$ and $\boldsymbol{T}$ to form a diagonalizable matrix whose eigenvectors are the parameters.
- Can write

$$\boldsymbol{S} \;=\; \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^\top$$

  where $\boldsymbol{W} = \mathrm{diag}(w_1, w_2, \ldots, w_k)$.
- What about $\boldsymbol{T}$?

"Flattening" a tensor to a matrix

- Recall multilinear function $\boldsymbol{T}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \sum_{i,j,k} T_{i,j,k} u_i v_j w_k$.
- Can also think of $\boldsymbol{T} \colon \mathbb{R}^d \to \mathbb{R}^{d \times d}$: (j,k)-th entry of $\boldsymbol{T}(\boldsymbol{u})$ is

$$\boldsymbol{T}(\boldsymbol{u})_{j,k} \;=\; \boldsymbol{T}(\boldsymbol{u}, \boldsymbol{e}_j, \boldsymbol{e}_k) \,.$$

  (Like "currying" in functional programming.)
- Since $\boldsymbol{T} = \sum_{t=1}^{k} w_t \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t \otimes \boldsymbol{z}_t$,

$$\begin{aligned} \boldsymbol{T}(\boldsymbol{u}) \;&=\; \sum_{t=1}^{k} w_t \langle \boldsymbol{z}_t, \boldsymbol{u} \rangle \cdot \boldsymbol{z}_t \otimes \boldsymbol{z}_t \\ &=\; \boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{u}}\boldsymbol{W}\boldsymbol{Z}^\top \end{aligned}$$

  where $\boldsymbol{D}_{\boldsymbol{u}} = \mathrm{diag}(\langle \boldsymbol{z}_1, \boldsymbol{u} \rangle, \langle \boldsymbol{z}_2, \boldsymbol{u} \rangle, \ldots, \langle \boldsymbol{z}_k, \boldsymbol{u} \rangle)$.

Combining the second and third moments

- Notation: $\boldsymbol{A}^\dagger$ is Moore-Penrose pseudoinverse of $\boldsymbol{A}$.
- Pick vector $\boldsymbol{u}$ somehow, and form $\boldsymbol{T}(\boldsymbol{u})\boldsymbol{S}^\dagger$.
  - $\boldsymbol{T}(\boldsymbol{u})\boldsymbol{S}^\dagger = (\boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{u}}\boldsymbol{W}\boldsymbol{Z}^\top)(\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^\top)^\dagger = \boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{u}}\boldsymbol{Z}^\dagger$
- Extract eigenvectors of $\boldsymbol{T}(\boldsymbol{u})\boldsymbol{S}^\dagger$ with non-zero eigenvalues.
  - If $\langle \boldsymbol{z}_t, \boldsymbol{u} \rangle \neq 0$, then $\boldsymbol{z}_t$ is eigenvector with non-zero eigenvalue.
  - **Caveat**: can only get $\boldsymbol{z}_t$ up to scaling $\sigma_t \boldsymbol{z}_t$ for $\sigma_t \neq 0$; but corresponding eigenvalue is $\langle \boldsymbol{z}_t, \boldsymbol{u} \rangle$, so

$$\sigma_t \;=\; \frac{\langle \sigma_t \boldsymbol{z}_t, \boldsymbol{u} \rangle}{\langle \boldsymbol{z}_t, \boldsymbol{u} \rangle}\,.$$

- Eigendecomposition is unique (up to scaling factors) as long as eigenvalues are distinct.
  - This holds if, e.g., $\boldsymbol{u}$ chosen uniformly at random from S^{d-1}.

Examples

Simple topic model

- Bag-of-words model for documents.
  - Document: seq. of tokens $X^{(1)}, X^{(2)}, \ldots$ from $\{1, 2, \ldots, d\}$.
  - Model parameters: $\{(w_t, \boldsymbol{\mu}_t)\}_{t=1}^k$, where $(w_1, w_2, \ldots, w_k) \in \Delta^{k-1}$ and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k \in \Delta^{d-1}$.
- Generative process:
  - Pick topic Y from $\{1, 2, \ldots, k\}$ with

$$\mathbb{P}(Y = t) \;=\; w_t \,.$$

  - Given Y, draw tokens iid from distribution $\boldsymbol{\mu}_Y$.
  - One-hot encode $X^{(i)} = j$ as vector $\boldsymbol{X}^{(i)} = \boldsymbol{e}_j \in \mathbb{R}^d$.
  - **Claim**:

$$\mathbb{E}(\boldsymbol{X}^{(1)} \otimes \boldsymbol{X}^{(2)}) \;=\; \sum_{t=1}^k w_t \cdot \boldsymbol{\mu}_t^{\otimes 2} \,,$$

$$\mathbb{E}(\boldsymbol{X}^{(1)} \otimes \boldsymbol{X}^{(2)} \otimes \boldsymbol{X}^{(3)}) \;=\; \sum_{t=1}^k w_t \cdot \boldsymbol{\mu}_t^{\otimes 3} \,.$$

Simple topic model: third-order moments

$$\boldsymbol{T} \;=\; \mathbb{E}\Big(\boldsymbol{X}^{(1)} \otimes \boldsymbol{X}^{(2)} \otimes \boldsymbol{X}^{(3)}\Big)$$

$$\begin{aligned} T_{i,j,k} \;&=\; \mathbb{P}(X^{(1)} = i \wedge X^{(2)} = j \wedge X^{(3)} = k) \\ &=\; \sum_t w_t \cdot \mathbb{P}(X^{(1)} = i \wedge X^{(2)} = j \wedge X^{(3)} = k \mid Y = t) \\ &=\; \sum_t w_t \cdot \mu_{t,i} \cdot \mu_{t,j} \cdot \mu_{t,k} \end{aligned}$$

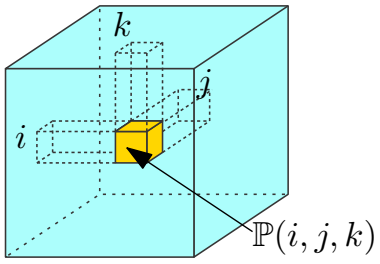

Figure 1: Third-order moments

Mixtures of spherical Gaussians

- Model parameters: $\{(w_t, \boldsymbol{\mu}_t, \sigma_i^2)\}_{t=1}^k$
  - $(w_1, w_2, \ldots, w_k) \in \Delta^{k-1}$;
  - $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$;
  - $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2 > 0$.
- For simplicity, assume $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$.
- $\boldsymbol{X} = \boldsymbol{Y} + \sigma \boldsymbol{Z}$, where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent, and
  - $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{\mu}_t) = w_t$ for each $t \in \{1, 2, \ldots, k\}$,
  - $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$.

Mixtures of spherical Gaussians: moments

- $\mathbb{E}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y}) = \sum_{t=1}^k w_t \cdot \boldsymbol{\mu}_t$
- $\mathbb{E}(\boldsymbol{X}^{\otimes 2}) = \sum_{t=1}^k w_t \cdot \boldsymbol{\mu}_t^{\otimes 2} + \sigma^2 \boldsymbol{I}$
  - How to estimate σ^2? Can look at smallest eigenvalue of centered second-order moment (i.e., covariance matrix).
- $\mathbb{E}(\boldsymbol{X}^{\otimes 3}) = \sum_{t=1}^k w_t \cdot \boldsymbol{\mu}_t^{\otimes 3} + f(\sigma^2, \mathbb{E}(\boldsymbol{X}))$
- **Upshot**: can estimate $\sum_{t=1}^k w_t \cdot \boldsymbol{\mu}_t^{\otimes 2}$ and $\sum_{t=1}^k w_t \cdot \boldsymbol{\mu}_t^{\otimes 3}$ from observable quantities.

Poisson topic model (following Canny, 2004)

- Bag-of-paragraphs model for documents.
  - Document: sequence of paragraphs (each a bag-of-words) $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots$
- Generative process:
  - Pick topic affinities $\boldsymbol{H} = (H_1, H_2, \ldots, H_k)$ from some non-Gaussian product distribution on $\mathbb{R}^k_{\geq 0}$.
  - Given $\boldsymbol{H}$, draw iid paragraph word count vectors $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots$ where

$$\boldsymbol{X}^{(i)} \mid \boldsymbol{H} \;\sim\; \bigotimes_{j=1}^{d} \mathrm{Poi}(\lambda_j)$$

    and

$$\boldsymbol{\lambda} \;=\; (\lambda_1, \lambda_2, \ldots, \lambda_d) \;=\; \sum_{t=1}^{k} H_t \boldsymbol{\mu}_t \,.$$

  - (Similar to Independent Component Analysis and Latent Dirichlet Allocation.)

Poisson topic model: moments

- First moment: $\boldsymbol{\mu} := \mathbb{E}(\boldsymbol{X}^{(1)}) = \sum_{t=1}^{k} \mathbb{E}(H_t) \cdot \boldsymbol{\mu}_t$.
- Second moment:

$$\begin{aligned}
&\mathbb{E}\Big[(\boldsymbol{X}^{(1)} - \boldsymbol{\mu}) \otimes (\boldsymbol{X}^{(2)} - \boldsymbol{\mu})\Big] \\
&\quad= \mathbb{E}\Big[\mathbb{E}\Big[(\boldsymbol{X}^{(1)} - \boldsymbol{\mu}) \otimes (\boldsymbol{X}^{(2)} - \boldsymbol{\mu}) \mid \boldsymbol{H}\Big]\Big] \\
&\quad= \mathbb{E}\left[\left(\sum_{t=1}^{k} (H_t - \mathbb{E}(H_t)) \cdot \boldsymbol{\mu}_t\right)^{\otimes 2}\right] \\
&\quad= \sum_{t=1}^{k} \mathrm{var}(H_t) \cdot \boldsymbol{\mu}_t^{\otimes 2} \,.
\end{aligned}$$

- Third moment:

$$\mathbb{E}\Big[(\boldsymbol{X}^{(1)} - \boldsymbol{\mu}) \otimes (\boldsymbol{X}^{(2)} - \boldsymbol{\mu}) \otimes (\boldsymbol{X}^{(3)} - \boldsymbol{\mu})\Big] \;=\; \sum_{t=1}^{k} \mathrm{skew}(H_t) \cdot \boldsymbol{\mu}_t^{\otimes 3} \,.$$

Multiview models

- Both topic models are examples of *multiview models*.
- Observables include multiple "views" $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}, \ldots$ that are conditionally independent given a hidden state H.
- **Example: hidden Markov model**
  - Hidden state sequence forms Markov chain (on finite state space): $H^{(1)} \to H^{(2)} \to H^{(3)} \to \cdots$
  - Observation sequence $X^{(1)}, X^{(2)}, X^{(3)}, \ldots$ where $X^{(t)}$ is independent of all other variables conditional on $H^{(t)}$.
  - **Multiview structure**: observables $X^{(1)}, X^{(2)}, X^{(3)}$ are conditionally independent given $H^{(2)}$.

Multiview moments

- View-i conditional means $\{\boldsymbol{\mu}_t^{(i)}\}_{t=1}^k$,

$$\boldsymbol{\mu}_t^{(i)} := \mathbb{E}\Big[\boldsymbol{X}^{(i)} \mid H = t\Big].$$

- Moments:

$$\mathbb{E}\Big[\boldsymbol{X}^{(1)} \otimes \boldsymbol{X}^{(2)} \otimes \boldsymbol{X}^{(3)}\Big] = \sum_{t=1}^{k} w_t \cdot \boldsymbol{\mu}_t^{(1)} \otimes \boldsymbol{\mu}_t^{(2)} \otimes \boldsymbol{\mu}_t^{(3)}$$

  where $w_t = \mathbb{P}(H = t)$.
- Not symmetric
  - Can be made symmetric using second-order moments.
  - Or, use asymmetric version of Jennrich's algorithm.