Higher-order moments and tensors

Daniel Hsu

COMS 4772

Higher-order moments

Support recovery from moments

- Random variable X, supported on k distinct points $z_1, z_2, \ldots, z_k \in \mathbb{R}$,

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

- How to learn parameters $\{(w_t, z_t)\}_{t=1}^{k}$?
  - Relatively straightforward given iid sample.
- **Statistical query model**:
  - Don't have iid sample, but instead can get (or estimate) $\mathbb{E}(f(X))$ for some simple functions f, e.g., $f(x) = x^2$.
  - Can we still learn parameters? What f to use?
  - **Prony's method**: uses functions $f(x) = x^p$ for $p \in \mathbb{N}$.

Prony's method (1795)

- $\mu_p := \mathbb{E}(X^p) = \sum_{t=1}^{k} w_t z_t^p$ (p-th moment of X)
- Use moments up to order $p = 2k - 1$.
  - There are $2k - 1$ parameters to estimate.
- Arrange into $k \times k$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$, where

$$G_{i,j} \; := \; \mu_{i+j-2} \,, \quad H_{i,j} \; := \; \mu_{i+j-1} \,;$$

called "Hankel matrices".

Hankel matrices of moments

$$\boldsymbol{G} := \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{k-1} \\ \mu_1 & \mu_2 & \cdots & \mu_k \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{k-1} & \mu_k & \cdots & \mu_{2k-2} \end{bmatrix}, \quad \boldsymbol{H} := \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_k \\ \mu_2 & \mu_3 & \cdots & \mu_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_k & \mu_{k+1} & \cdots & \mu_{2k-1} \end{bmatrix}.$$

Matrix factorization of $\boldsymbol{G}$

$$G_{i,j} = \mu_{i+j-2} = \sum_{t=1}^{k} w_t z_t^{i+j-2} = \sum_{t=1}^{k} w_t z_t^{i-1} z_t^{j-1}$$

Therefore

$$\boldsymbol{G} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{V}^\top$$

where

$$\boldsymbol{W} := \operatorname{diag}(w_1, w_2, \ldots, w_k) \succ 0$$

and $\boldsymbol{V}$ is a Vandermonde matrix

$$V_{i,t} := z_t^{i-1}$$

whose determinant is

$$\det(\boldsymbol{V}) = \prod_{1 \le s < t \le k} (z_s - z_t) \neq 0.$$

Matrix factorization of $\boldsymbol{H}$

$$H_{i,j} \;=\; \mu_{i+j-1} \;=\; \sum_{t=1}^{k} w_t z_t^{i+j-1} \;=\; \sum_{t=1}^{k} (z_t w_t) z_t^{i-1} z_t^{j-1}$$

Therefore

$$\boldsymbol{H} = \boldsymbol{V} \boldsymbol{Z} \boldsymbol{W} \boldsymbol{V}^{\top}$$

where

$$\boldsymbol{Z} \;:=\; \operatorname{diag}(z_1, z_2, \ldots, z_k)$$

and $\boldsymbol{W}$ and $\boldsymbol{V}$ are as before.

Prony's method (finally)

- (Not exactly Prony's method, but similar.)
- Form $\boldsymbol{G}$ and $\boldsymbol{H}$ using moments of orders $0 \le p \le 2k - 1$.
  - Recall $\boldsymbol{G} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{V}^{\top}$ and $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{V}^{\top}$.
- Compute $\boldsymbol{H}\boldsymbol{G}^{-1}$.
  - $\boldsymbol{H}\boldsymbol{G}^{-1} = (\boldsymbol{V}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{V}^{\top})(\boldsymbol{V}\boldsymbol{W}\boldsymbol{V}^{\top})^{-1} = \boldsymbol{V}\boldsymbol{Z}\boldsymbol{V}^{-1}$.
- Compute eigenvalues of $\boldsymbol{H}\boldsymbol{G}^{-1}$.
  - $\boldsymbol{V}\boldsymbol{Z}\boldsymbol{V}^{-1}$ is diagonalizable; eigenvalues are $z_1, z_2, \ldots, z_k$.
  - Get $\{z_t\}_{t=1}^{k}$ (in some arbitrary order).
- Form $\boldsymbol{V}$ (up to permutation of columns); compute $\boldsymbol{V}^{-1}\boldsymbol{G}\boldsymbol{V}^{-\top}$.
  - This equals $\boldsymbol{W}$ (up to *same* permutation).
  - Can read out $\{w_t\}_{t=1}^{k}$ from diagonal entries, match with z_t's.

Learning convex polytopes (Gravin et al, 2011)

- Convex polytope $P \subset \mathbb{R}^d$ with vertices $Z := \{\boldsymbol{z}_t\}_{t=1}^k \in \mathbb{R}^d$.
- Assume P is *simple*:
  - Every vertex $\boldsymbol{z}_t$ is incident on d edges, i.e., has d neighboring vertices $\boldsymbol{z}_t^{(1)}, \boldsymbol{z}_t^{(2)}, \dots, \boldsymbol{z}_t^{(d)} \in Z$, such that $\{\boldsymbol{z}_t^{(i)} - \boldsymbol{z}_t\}_{i=1}^d$ are linearly independent.
- Suppose $\boldsymbol{X} \sim \text{Uniform}(P)$. Can we learn the vertices Z?

Algorithm based on Prony's method

- For any $\boldsymbol{u} \in S^{d-1}$,

$$\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{X} \rangle^p \;=\; c_{p,d} \cdot \sum_{t=1}^k D_t(\boldsymbol{u}) \cdot \langle \boldsymbol{u}, \boldsymbol{z}_t \rangle^{p+d}$$

  where $c_{p,d} > 0$ is some constant depending only on p and d,

$$D_t(\boldsymbol{u}) \;=\; \frac{|\det(\boldsymbol{K}_t)|}{\prod_{i=1}^d \langle \boldsymbol{u}, \boldsymbol{z}_t^{(i)} - \boldsymbol{z}_t \rangle},$$

  and

$$\boldsymbol{K}_t \;=\; \left[\boldsymbol{z}_t^{(1)} - \boldsymbol{z}_t \mid \boldsymbol{z}_t^{(2)} - \boldsymbol{z}_t \mid \cdots \mid \boldsymbol{z}_t^{(d)} - \boldsymbol{z}_t \right].$$

- Can use Prony's method to learn $\{\langle \boldsymbol{u}, \boldsymbol{z}_t \rangle\}_{t=1}^k$.
- Do this for d different $\boldsymbol{u}$'s to reconstruct $\{\boldsymbol{z}_t\}_{t=1}^k$.

Higher-order moments

- Moments provide constraints on parameters:
  - **Prony's method**: use moments up to order $2k-1$ to solve for $2k-1$ unknowns.
  - **Convex polytope learning**: use moments up to order $2k-1$ (d times) to solve for dk unknowns.
- Variance of high-order moments is high – can be difficult to estimate accurately.
- But typically $\boldsymbol{X}$ in $\mathbb{R}^d$ has $\Omega(d^p)$ *mixed* moments of order p.
  - E.g., $\mathbb{E}(X_1^2 X_5)$.
  - Perhaps we can get away with moments of small order?

Multilinear functions and tensors

Motivation: Spearman's hypothesis

- **Spearman's hypothesis**: a student's test score depends on
  - how much test measures *math* and *verbal* abilities;
  - student's abilities in *math* and *verbal*.
- Model: score for student i on test j given by

$$S(i,j) \ := \ x_{\mathrm{math}}(i) \cdot y_{\mathrm{math}}(j) + x_{\mathrm{verbal}}(i) \cdot y_{\mathrm{verbal}}(j) \, .$$

  - $x_{\mathrm{math}}(i)$ and $x_{\mathrm{verbal}}(i)$ are math and verbal abilities of student i
  - $y_{\mathrm{math}}(j)$ and $y_{\mathrm{verbal}}(j)$ are math-iness and verbal-iness of test j
- Matrix equation ($\boldsymbol{X} = [\boldsymbol{x}_{\mathrm{math}} \mid \boldsymbol{x}_{\mathrm{verbal}}]$, $\boldsymbol{Y} = [\boldsymbol{y}_{\mathrm{math}} \mid \boldsymbol{y}_{\mathrm{verbal}}]$):

$$\boldsymbol{S} \ = \ \boldsymbol{X}\boldsymbol{Y}^{\top} \, .$$

- But why "math" and "verbal"?

$$\boldsymbol{S} \ = \ (\boldsymbol{X}\boldsymbol{R})(\boldsymbol{Y}\boldsymbol{R}^{-\top})^{\top}$$

  for any 2×2 invertible matrix $\boldsymbol{R}$.

Matrices

- Matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ as *bilinear function* $\boldsymbol{M} \colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$.
- Linear in each argument:
  $\boldsymbol{M}(c\boldsymbol{x} + \boldsymbol{x}', \boldsymbol{y}) = c\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}', \boldsymbol{y})$
  $\boldsymbol{M}(\boldsymbol{x}, c\boldsymbol{y} + \boldsymbol{y}') = c\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}')$
- Formula using matrix represetation: $\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^{\top}\boldsymbol{M}\boldsymbol{y}$.
- Using singular value decomposition $\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_j^{\top}$:
  $\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{r} \sigma_i \langle \boldsymbol{u}_i, \boldsymbol{x} \rangle \langle \boldsymbol{v}_i, \boldsymbol{y} \rangle$
- Forget about matrix representation. How to describe $\boldsymbol{M}$?
  - Pick any bases $\{\boldsymbol{e}_i\}_{i=1}^{m}$ for $\mathbb{R}^m$ and $\{\boldsymbol{f}_j\}_{j=1}^{n}$ for $\mathbb{R}^n$
  - Describe $\boldsymbol{M}$ by $m \times n$ function values $\boldsymbol{M}(\boldsymbol{e}_i, \boldsymbol{f}_j)$.

p-linear functions

- $\boldsymbol{T} \colon \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_p} \to \mathbb{R}$
- Describe $\boldsymbol{T}$ by its behavior of basis elements, e.g., $\{\boldsymbol{e}_i^{(1)}\}_{i=1}^{n_1}, \dots, \{\boldsymbol{e}_i^{(p)}\}_{i=1}^{n_p}$:
$$\boldsymbol{T}(\boldsymbol{e}_{i_1}^{(1)}, \dots, \boldsymbol{e}_{i_p}^{(p)})$$
($n_1 \times n_2 \times \cdots \times n_p$ function values.)
- Like matrices, can arrange into multi-index array $\boldsymbol{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_p}$
- Also called a **p-th order tensor**
  - $p = 1$: vector in $\mathbb{R}^n$
  - $p = 2$: matrix in $\mathbb{R}^{m \times n}$
  - We will usually just consider $p = 3$ for simplicity
- "Formula" using multi-index array $\boldsymbol{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$:
$$\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \;=\; \sum_{i,j,k} T_{i,j,k} x_i y_j z_k \,.$$

Tensor norms

- Frobenius norm: $\|\boldsymbol{T}\|_F = \sqrt{\sum_{i,j,k} T_{i,j,k}^2}$
- Operator (spectral) norm: $\|\boldsymbol{T}\|_2 = \max_{\substack{\boldsymbol{x} \in S^{n_1 - 1}, \\ \boldsymbol{y} \in S^{n_2 - 1}, \\ \boldsymbol{z} \in S^{n_3 - 1}}} \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$
  - Optimization problem is NP-hard for $p \geq 3$.
  - In fact, **most problems we take for granted as tractable for matrices are NP-hard for tensors of order** $p \geq 3$.

Rank

- **Rank-1 tensor**:

$$\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \; = \; \langle \boldsymbol{u}, \boldsymbol{x} \rangle \langle \boldsymbol{v}, \boldsymbol{y} \rangle \langle \boldsymbol{w}, \boldsymbol{z} \rangle$$

  for some vectors $\boldsymbol{u} \in \mathbb{R}^{n_1}, \boldsymbol{v} \in \mathbb{R}^{n_2}, \boldsymbol{w} \in \mathbb{R}^{n_3}$.
  - Write as $\boldsymbol{T} = \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}$
  - Multi-index array: $T_{i,j,k} = u_i v_j w_k$
  - Generalization of matrix "outer product" $\boldsymbol{u}\boldsymbol{v}^\top \equiv \boldsymbol{u} \otimes \boldsymbol{v}$
- Say $\mathrm{rank}(\boldsymbol{T})$ = smallest $r \in \mathbb{N}$ such that $\boldsymbol{T}$ equals the sum of r rank-1 tensors.
  - Generalizes concept of matrix rank.
- Computing rank is NP-hard for $p \geq 3$.

Border rank

- **Border rank of $\boldsymbol{T}$**: smallest $r \in \mathbb{N}$ such that there exists a sequence $(\boldsymbol{T}_k)_{k \in \mathbb{N}}$ of rank r tensors such that $\lim_{k \to \infty} \boldsymbol{T}_k = \boldsymbol{T}$
- **In general, border rank not the same as rank.**
- Example:
  - Take any distinct $\boldsymbol{u}, \boldsymbol{v} \in S^{n-1}$, and define

$$\boldsymbol{T} \; := \; \boldsymbol{u} \otimes \boldsymbol{u} \otimes \boldsymbol{v} + \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{u} + \boldsymbol{v} \otimes \boldsymbol{u} \otimes \boldsymbol{u} \,,$$

    which has rank 3.
  - Define

$$\boldsymbol{T}_{1/\epsilon} \; := \; \frac{1}{\epsilon}(\boldsymbol{u} + \epsilon \boldsymbol{v}) \otimes (\boldsymbol{u} + \epsilon \boldsymbol{v}) \otimes (\boldsymbol{u} + \epsilon \boldsymbol{v}) - \frac{1}{\epsilon} \boldsymbol{u} \otimes \boldsymbol{u} \otimes \boldsymbol{u} \,.$$

  - For $\epsilon = 1/k$, have $\lim_{k \to \infty} \boldsymbol{T}_k = \boldsymbol{T}$.

Uniqueness of decompositions

- Suppose $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$ are orthonormal.
- Matrix: $\boldsymbol{M} = \sum_{i=1}^{n} \boldsymbol{v}_i \otimes \boldsymbol{v}_i = \sum_{i=1}^{n} \boldsymbol{v}_i^{\otimes 2}$.
  - **Cannot recover $\{\boldsymbol{v}_i\}_{i=1}^{n}$ just from $\boldsymbol{M}$.**
- 3rd-order tensor: $\boldsymbol{T} = \sum_{i=1}^{n} \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i = \sum_{i=1}^{n} \boldsymbol{v}_i^{\otimes 3}$.
  - ***Can* recover $\{\boldsymbol{v}_i\}_{i=1}^{n}$ just from $\boldsymbol{T}$ exactly!**
- Many general conditions imply uniqueness of higher-order tensor decomposition.