Optimization for Tensor Models

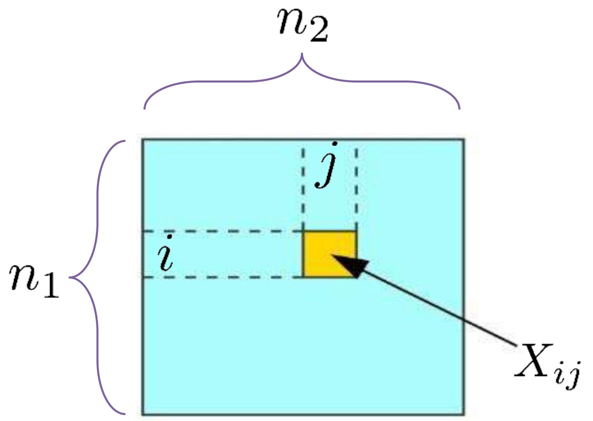
Donald Goldfarb
IEOR Department
Columbia University

UCLA Mathematics Department
Distinguished Lecture Series
May 17 – 19, 2016

Tensors

Matrix

$$\boldsymbol{X} = (X_{ij}) \in \mathbb{R}^{n_1 \times n_2}$$



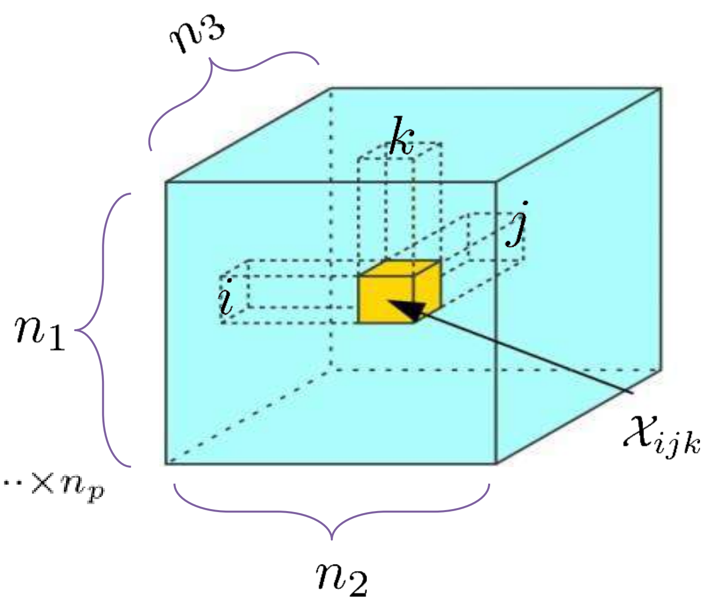
Tensor: higher-order matrix

three-way tensor:

$$\boldsymbol{\mathcal{X}} = (\mathcal{X}_{ijk}) \in \mathbb{R}^{n_1 \times n_2 \times n_3}$$

p-way tensor:

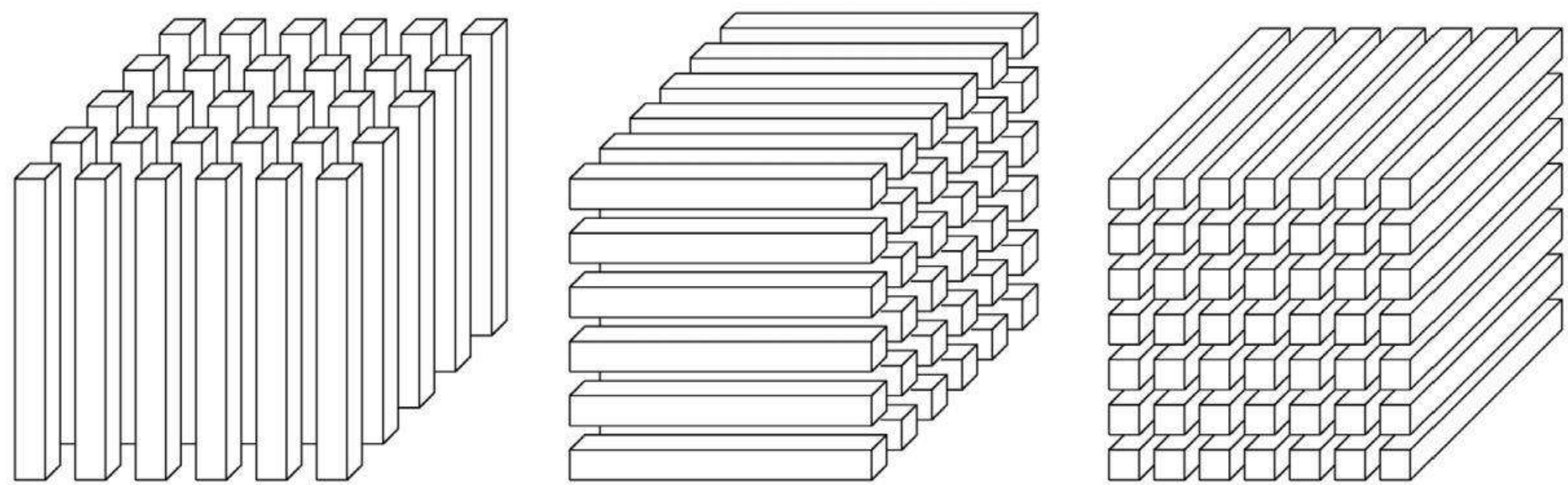
$$\boldsymbol{\mathcal{X}} \in \bigotimes_{i=1}^{p} \mathbb{R}^{n_i} := (\mathcal{X}_{i_1 i_2 \cdots i_k}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_p}$$



Tensors

Tensor == Vector ? Fibers?

can be viewed as a collection of fibers



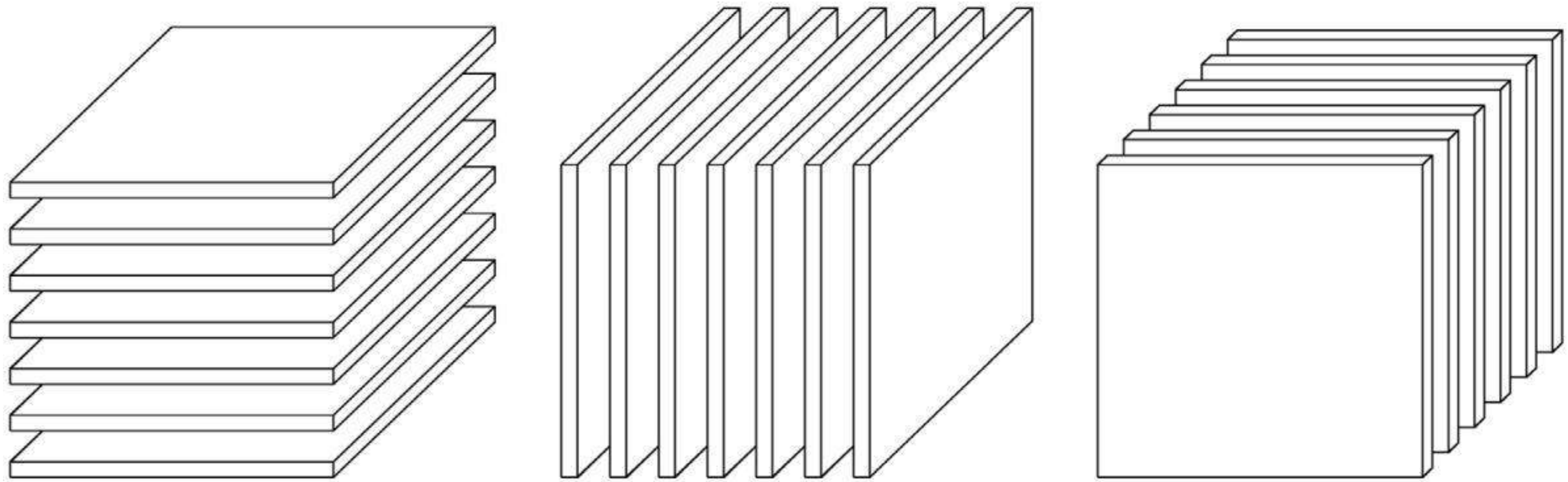
(a) Mode-1 (column) fibers: $\mathbf{x}_{:jk}$ (b) Mode-2 (row) fibers: $\mathbf{x}_{i:k}$ (c) Mode-3 (tube) fibers: $\mathbf{x}_{ij:}$

(Kolda & Bader, 2009)

Tensors

Tensor == Matrix ? Slices?

can be viewed as a collection of slices



(a) Horizontal slices: $\mathbf{X}_{i::}$ (b) Lateral slices: $\mathbf{X}_{:j:}$ (c) Frontal slices: $\mathbf{X}_{::k}$ (or $\mathbf{X}_k$)

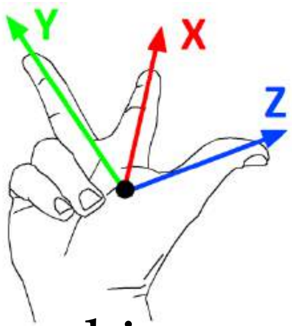
(Kolda & Bader, 2009)

Tensors

Why tensors?

- tensors capture multilinear structure
- more **flexible** and **powerful** models

  e.g. parameter estimation in latent variable modelling
  (to be discussed shortly)



Tensor: object in its own right

- its own geometrical, statistical and computational issues
- much harder to work with than a matrix

Example: Single Topic Model

Model:

- k topics (dists. over d words)

  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$

- sample topic h:

  prob$[h = i] = w_i$ $(i \in [k])$

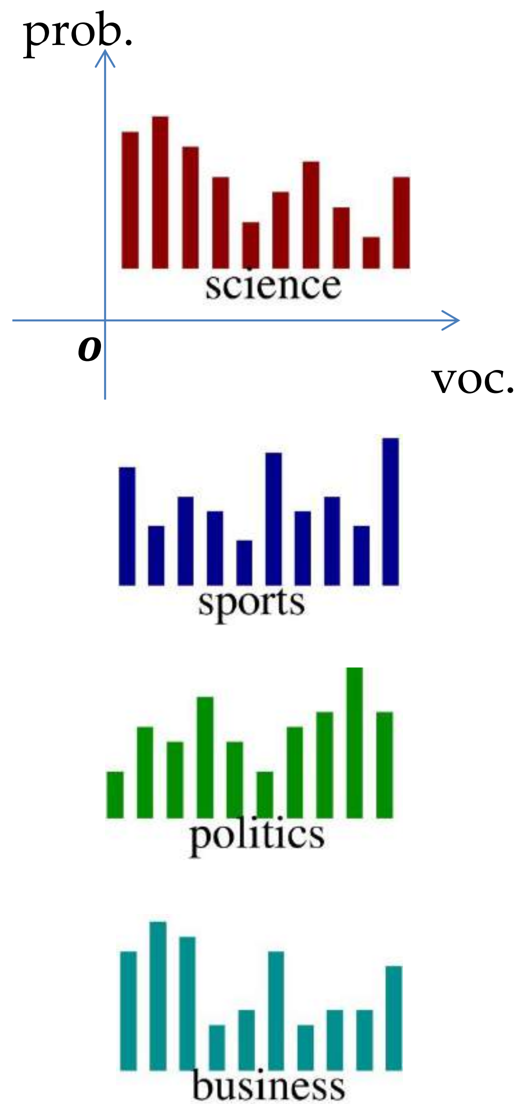
- each document has $\boldsymbol{m}$ words

  $x_1, x_2, \ldots, x_m$ sampled i.i.d. from $\boldsymbol{\mu}_i$

Dataset: m-word documents

Goal: learn parameters

$$\boldsymbol{\theta} = \left(\{\boldsymbol{\mu}_i\}_{i \in [k]}, \{w_i\}_{i \in [k]} \right)$$



Example: Single Topic Model

Method of moments:

Karl Pearson (1857~1936)

Key idea: find parameters (approx.) consistent with observed moments (Pearson, 1894)

Procedure:

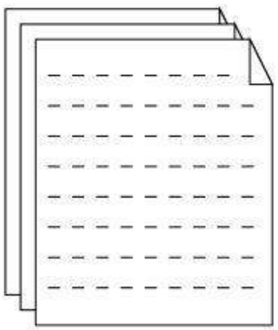
- setup equations

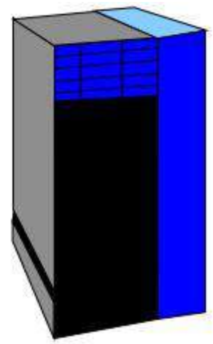
$$\boldsymbol{Moment}_{population}\ (\boldsymbol{\theta}) = \widehat{\boldsymbol{Moment}}_{sample}$$

- solve approximately

Example: Single Topic Model

Goal: model the topics of the documents in a given corpus





$\boldsymbol{\theta}$

Samples of documents generated based on single topic model with params $\boldsymbol{\theta} = (\{\boldsymbol{\mu_i}\}, \{w_i\})$

Unsupervised learning alg.
Method of moments
- ☐ setup equations
- ☐ solve them approx.

Model parameters

Which moments to use?
1st -order? 2nd -order? 3rd -order? pth -order?

Example: Single Topic Model

Binary encoding:

$\boldsymbol{x}_t = \boldsymbol{e}_i \Longleftrightarrow$ the t-th word in the document is i-th word in the vocabulary

$(t, i) \in [m] \times [d]$

E.g.

topic: animal

3-word document

7-word vocabulary

vocab.	x_1	x_2	x_3
cats	0	0	1
dogs	0	0	0
fear	0	0	0
I	1	0	0
like	0	1	0
raise	0	0	0
want	0	0	0

Example: Single Topic Model

Binary encoding:

$\boldsymbol{x}_t = \boldsymbol{e}_i \iff$ the t-th word in the document is i-th word in the vocabulary

$(t, i) \in [m] \times [d]$

First moment

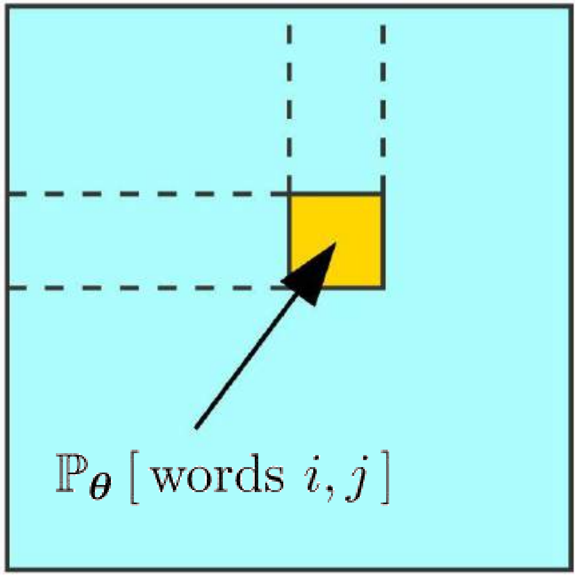
$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_1] &= \left(\mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{x}_1 = \boldsymbol{e}_i]\right)_{i \in [d]} \\
&= \mathbb{E}_h[\mathbb{E}[\boldsymbol{x}_1 \mid \text{topic } h]] \\
&= \mathbb{E}_h[\boldsymbol{\mu}_h] \\
&= \sum_{i \in [k]} w_i \boldsymbol{\mu}_i \in \mathbb{R}^d
\end{aligned}$$

Not identifiable: only d numbers for $(d+1)k$ parameters.

Example: Single Topic Model

Second moment

$$
\begin{aligned}
\boldsymbol{M} &= \mathbb{E}\left[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2\right] \\
&= \left(\mathbb{P}_{\boldsymbol{\theta}}\left[\,(\boldsymbol{x}_1, \boldsymbol{x}_2) = (i, j)\,\right]\right)_{(i,j)\in[d]\times[d]} \\
&= \mathbb{E}_h\left[\mathbb{E}\left[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \mid \text{topic } h\right]\right] \\
&= \mathbb{E}_h\left[\boldsymbol{\mu}_h \otimes \boldsymbol{\mu}_h \mid \text{topic } h\right] \\
&= \sum_{i\in[k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \in \mathbb{R}^{d\times d}
\end{aligned}
$$



Matrix-mode: still not identifiable even though $\frac{d(d+1)}{2} > (d+1)k$. (Why?)

Example: Single Topic Model

Identifiable? NO!

$\boldsymbol{U} \in \mathbb{R}^{n\times k}$ is a solution

$\Longleftrightarrow$ $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{U}^T \in \mathbb{R}^{n\times n}$

$\Longleftrightarrow$ $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{Q}(\boldsymbol{U}\boldsymbol{Q})^T$, for any $\boldsymbol{Q} \in \mathcal{O}_k$ $(\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_k)$

$\Longleftrightarrow$ $\boldsymbol{U} \in \mathbb{R}^{n\times k}$ is a solution for any $\boldsymbol{Q} \in \mathcal{O}(k)$

$\Longleftrightarrow$ AMBUIGUITY!

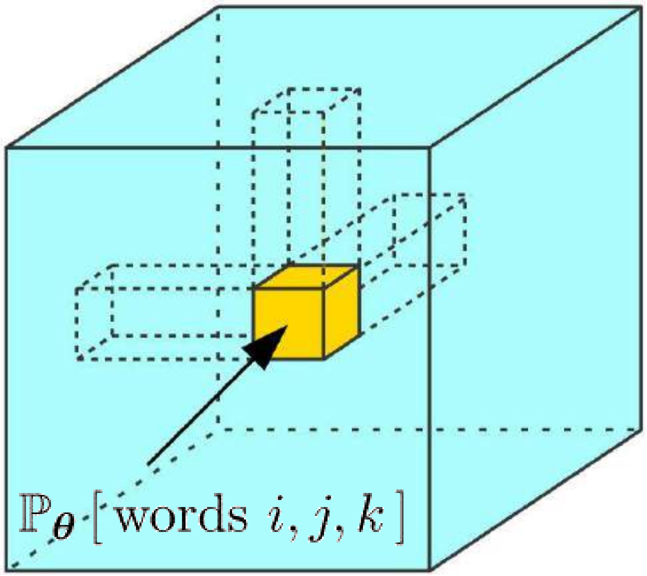
Matrix mode: INSUFFICIENT to identify parameters.!

Example: Single Topic Model

Third-order moment

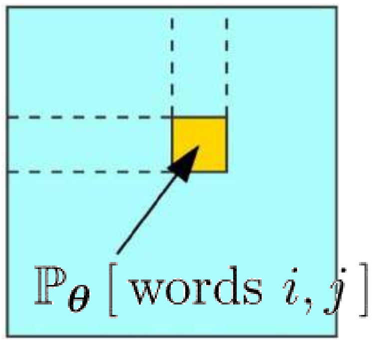
$$
\begin{aligned}
\boldsymbol{\mathcal{T}} &= \mathbb{E}\left[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3\right] \\
&= \Big(\mathbb{P}_{\boldsymbol{\theta}}\left[\,(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = (i, j, k)\,\right]\Big)_{(i,j,k)\in[d]\times[d]\times[d]} \\
&= \mathbb{E}_h\left[\mathbb{E}\left[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3 \mid \text{topic } h\right]\right] \\
&= \mathbb{E}_h\left[\boldsymbol{\mu}_h \otimes \boldsymbol{\mu}_h \otimes \boldsymbol{\mu}_h \mid \text{topic } h\right] \\
&= \sum_{i\in[k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \in \mathbb{R}^{d\times d\times d}
\end{aligned}
$$

Tensor mode

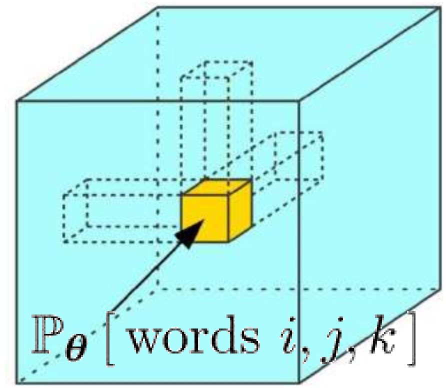


Example: Single Topic Model

$M_{\boldsymbol{\theta}} :=$



$\mathcal{T}_{\boldsymbol{\theta}} :=$



$$M_{\boldsymbol{\theta}} = \sum_{i \in [k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \qquad \mathcal{T}_{\boldsymbol{\theta}} = \sum_{i \in [k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

Claim: $M_{\boldsymbol{\theta}}$ & $\mathcal{T}_{\boldsymbol{\theta}}$ **uniquely determine the parameters** $\boldsymbol{\theta}$

Example: Single Topic Model

Reduction to orthogonal case via whitening

Whiten:

project to k dimensions

transform to orthogonality

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$$

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{D}^{-1/2}$$

Reduced Eigen-Decomp.

$$\boldsymbol{M} = \sum_{i\in[k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$\mathcal{T} = \sum_{i\in[k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

apply $\boldsymbol{W}$ to $\boldsymbol{M}$ and $\boldsymbol{T}$

$$\boldsymbol{v}_i = \sqrt{w_i}\boldsymbol{W}^\top \boldsymbol{\mu}_i \in \mathbb{R}^k$$

$$\lambda_i = 1/\sqrt{w_i}$$

$$\widetilde{\boldsymbol{M}} = \sum_{i\in[k]} \boldsymbol{v}_i \otimes \boldsymbol{v}_i = \boldsymbol{I}_k$$

$$\widetilde{\mathcal{T}} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i$$

$\{\boldsymbol{v}_\mathrm{i}\}_{\mathrm{i}\in[k]}$ forms an orthonormal basis for $\mathbb{R}^k$

Example: Single Topic Model

Spectral theorem and eigen-decompositions

$\{\boldsymbol{v}_{\mathrm{i}}\}_{\mathrm{i}\in[k]}$ forms an orthnormal basis for $\mathbb{R}^k$

symmetric matrix

$$\boldsymbol{X} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i$$

eigen-decomp. is unique iff
$\lambda_i \neq \lambda_j$

symmetric tensor

$$\boldsymbol{\mathcal{X}} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i$$

if such decomp. exists,
then it is always unique
(even if λ_i's all same)

Uniqueness of orthogonal decomp. → $\boldsymbol{M}_{\boldsymbol{\theta}}$ & $\boldsymbol{\mathcal{T}}_{\boldsymbol{\theta}}$ uniquely determine the parameters $\boldsymbol{\theta}$

Identifiability issue is resolved via tensor mode!

2^{nd} Example: Mixture of Spherical Gaussians

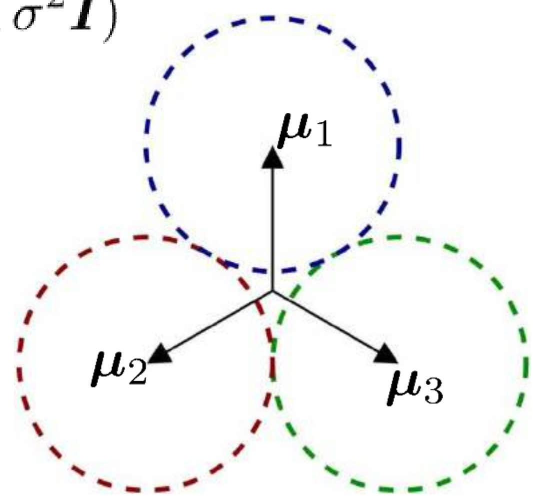
Model:

k means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$

Sample cluster $h = i$ with probability w_i $(i \in [k])$

Observe $\boldsymbol{x}$, with i.i.d. homogeneous spherical noise

$$\boldsymbol{x} = \boldsymbol{\mu}_i + \boldsymbol{\eta}, \quad \eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$



Dataset: multiple points

Goal: learn parameters

$$\boldsymbol{\theta} = \left(\{\boldsymbol{\mu}_i\}_{i\in[k]}, \{w_i\}_{i\in[k]}, \sigma \right)$$

2nd Example: Mixture of Spherical Gaussians

Identifiable using 1st, 2nd and 3rd –order moments together

[Hsu & Kakade, '13]

$$\sigma^2 = \lambda_{\min}(\mathbb{E}\left[\boldsymbol{x} \otimes \boldsymbol{x}\right])$$

$$\boldsymbol{M} = \mathbb{E}\left[\boldsymbol{x} \otimes \boldsymbol{x}\right] - \sigma^2 \boldsymbol{I} = \sum_{i \in [k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

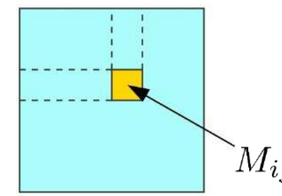
$$\begin{aligned} \mathcal{T} &= \mathbb{E}\left[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}\right] \\ &\qquad - \sigma^2 \sum_{i \in [d]} \left(\mathbb{E}\left[\boldsymbol{x}\right] \otimes \boldsymbol{e}_i \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \mathbb{E}\left[\boldsymbol{x}\right] \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \boldsymbol{e}_i \otimes \mathbb{E}\left[\boldsymbol{x}\right]\right) \\ &= \sum_{i \in [k]} w_i \mu_i \otimes \mu_i \otimes \mu_i \end{aligned}$$

Same approach as simple topic model!

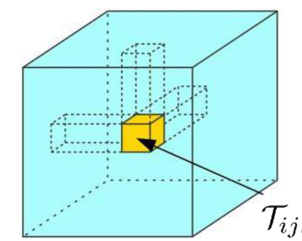
General Principle

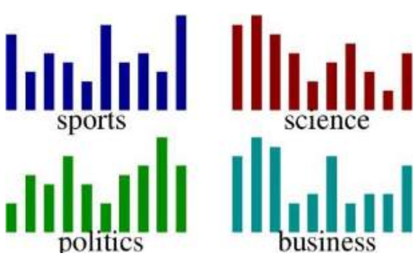
Similar structures prevail in latent variable models

$$\boldsymbol{M} = f\Big(\ \leq \text{2nd-order moments} \ \Big) = \sum_{i\in[k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

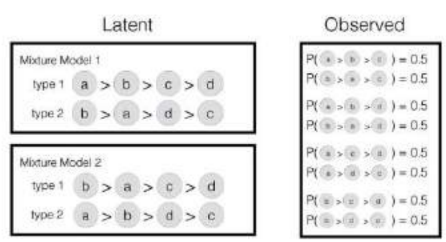


$$\boldsymbol{\mathcal{T}} = g\Big(\ \leq \text{3rd-order moments} \ \Big) = \sum_{i\in[k]} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

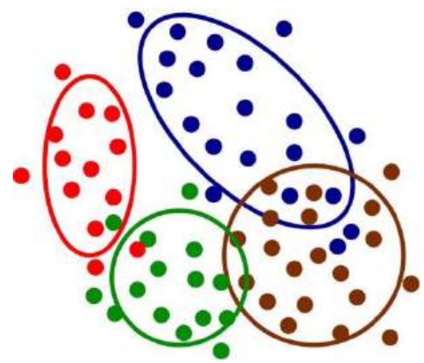




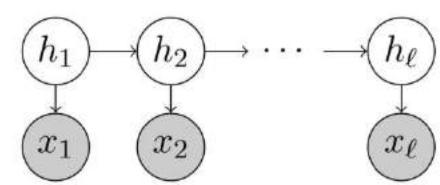
Latent Dirichlet Allocatioin (LDA)



Mixed Multinomial Logit Model



Mixture of Gaussians (MoG)



Hidden Markov Models (HMMs)

Orthogonal Decomposition

How to find $\{(\lambda_i, \boldsymbol{v}_\mathrm{i})\}_{\mathrm{i}\in[k]}$? $(<\boldsymbol{v}_\mathrm{i}, \boldsymbol{v}_\mathrm{j}> = \boldsymbol{\delta}_{ij})$

symmetric matrix

$$\boldsymbol{X} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i$$

successive rank-one approximation (SROA)

symmetric tensor

$$\boldsymbol{\mathcal{X}} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i$$

generalized SROA?

Successive Rank-One Approximation

Symmetric matrix

$$\boldsymbol{M} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \qquad (\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij})$$

SROA: rank-one approximation **+** deflation

Repeat k times

$$(\hat{\lambda}, \hat{\boldsymbol{v}}) \in \operatorname*{arg\,min}_{\lambda\in\mathbb{R}, \|\boldsymbol{v}\|=1} \|\boldsymbol{M} - \lambda \boldsymbol{v} \otimes \boldsymbol{v}\| \qquad \textbf{(rank-one approx.)}$$

$$\boldsymbol{M} \leftarrow \boldsymbol{M} - \hat{\lambda} \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \qquad \textbf{(deflation)}$$

Recover $\{(\lambda_i, \boldsymbol{v}_i)\}_{i\in[k]}$ exactly ($\lambda_i \neq \lambda_j$)

Successive Rank-One Approximation

Symmetric orthogonal decomposable (SOD) tensor

$$\mathcal{T} = \sum_{i \in [k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i \qquad (\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij})$$

SROA: rank-one approximation + deflation

Repeat k times

$$(\hat{\lambda}, \hat{\boldsymbol{v}}) \in \operatorname*{arg\,min}_{\lambda \in \mathbb{R}, \|\boldsymbol{v}\|=1} \|\mathcal{T} - \lambda \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v}\|_F \quad \textbf{(rank-one approx.)}$$

$$\mathcal{T} \leftarrow \mathcal{T} - \hat{\lambda} \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \quad \textbf{(deflation)}$$

Recover $\{(\lambda_i, \boldsymbol{v}_i)\}_{i \in [k]}$ exactly up to sign flips [Zhang & Golub, 01]

N.B.: no requirement on λ's to be distinct

Successive Rank-One Approximation

Sampling error

$$\#\text{ samples} = \infty \;\;\Rightarrow\;\; \hat{\mathcal{T}} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i$$

$$\#\text{ samples} < \infty \;\;\Rightarrow\;\; \hat{\mathcal{T}} = \sum_{i\in[k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i + \boldsymbol{\mathcal{E}}$$

Other potential sources of perturbation $\mathcal{E}$

- Model misspecification
- Numerical error
- ……

Is SROA **robust** to the perturbation?

Recall **matrix perturbation theory** (e.g. Davis-Kahan),
which requires $\|\, perturbation\ matrix \,\| < \min_{i\neq j} |\lambda_i - \lambda_j|.$

Successive Rank-One Approximation

Perturbed SOD tensor

$$\widehat{\mathcal{T}} = \sum_{i \in [k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i + \boldsymbol{\mathcal{E}} \qquad (\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij})$$

SROA: rank-one approximation + deflation

Repeat k times

$$(\hat{\lambda}, \hat{\boldsymbol{v}}) \in \operatorname*{arg\,min}_{\lambda \in \mathbb{R}, \|\boldsymbol{v}\|=1} \left\| \widehat{\mathcal{T}} - \lambda \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v} \right\|_F \quad \textbf{(rank-one approx.)}$$

$$\widehat{\mathcal{T}} \leftarrow \widehat{\mathcal{T}} - \hat{\lambda} \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \quad \textbf{(deflation)}$$

Successive Rank-One Approximation

Input: the perturbed SOD tensor

Repeat k times

$$(\hat{\lambda}, \hat{\boldsymbol{v}}) \in \operatorname*{arg\,min}_{\lambda \in \mathbb{R}, \|\boldsymbol{v}\|=1} \left\| \widehat{\mathcal{T}} - \lambda \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v} \right\|_F$$ **(rank-one approx.)**

$$\widehat{\mathcal{T}} \leftarrow \widehat{\mathcal{T}} - \hat{\lambda} \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}}$$ **(deflation)**

Output: $\left\{ (\hat{\lambda}_i, \hat{\boldsymbol{v}}_i) \right\}_{i \in [k]}$

Theorem (MHG, '15)

Output $\left\{ (\hat{\lambda}_i, \hat{\boldsymbol{v}}_i) \right\}_{i \in [k]}$

$\|\boldsymbol{\mathcal{E}}\| \leq C \cdot \dfrac{\lambda_{\min}}{\sqrt{n}}$

➔

$\exists$ perm. π on $[k]$ s.t. $\forall\, i \in [k]$

$$\begin{cases} \min_{\pm} \left\{ \left| \lambda_{\pi(i)} \pm \hat{\lambda}_i \right| \right\} \leq 2\varepsilon \\ \min_{\pm} \left\{ \left\| \boldsymbol{v}_{\pi(i)} \pm \hat{\boldsymbol{v}}_i \right\| \right\} \leq 20\varepsilon / \lambda_{\pi(i)} \end{cases}$$

N.B.

- generalizes matrix perturbation analysis
- **NO** spectral gap quantity involved

Best Rank-One Tensor Approximation

$$\text{minimize} \quad \|\boldsymbol{\mathcal{T}} - \lambda \underbrace{\boldsymbol{v} \otimes \boldsymbol{v} \otimes \cdots \otimes \boldsymbol{v}}_{p \text{ times}}\|_F^2 \quad \text{subject to} \quad \|\boldsymbol{v}\| = 1, \quad \lambda \in \mathbb{R}$$

$$\Updownarrow \qquad \|\boldsymbol{\mathcal{T}} - \lambda \boldsymbol{v}^{\otimes p}\|_F^2 = \|\boldsymbol{\mathcal{T}}\|_F^2 - 2\lambda \langle \boldsymbol{\mathcal{T}}, \boldsymbol{v}^{\otimes p} \rangle + \lambda^2$$

$$\min_{\|\boldsymbol{v}\|=1} \min_{\lambda \in \mathbb{R}} \left(-2\lambda \langle \boldsymbol{\mathcal{T}}, \boldsymbol{v}^{\otimes p} \rangle + \lambda^2 \right)$$

$\Updownarrow$ **eliminate** λ, then **solve**

$$\min_{\|\boldsymbol{v}\|=1} \langle \boldsymbol{\mathcal{T}}, \boldsymbol{v}^{\otimes p} \rangle \quad \& \quad \min_{\|\boldsymbol{v}\|=1} \langle -\boldsymbol{\mathcal{T}}, \boldsymbol{v}^{\otimes p} \rangle$$

SDP Relaxation [Jiang, Ma & Zhang, '14]

$$\text{minimize} \quad \langle \mathcal{T}, \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v} \rangle \quad \text{subject to} \quad \|\boldsymbol{v}\| = 1$$ (*)

$\Updownarrow$ **reshape** $\mathcal{T} \in \bigotimes^4 \mathbb{R}^n$ to $\boldsymbol{T}_\square \in \mathbb{R}^{n^2 \times n^2}$

$$\text{minimize} \quad \left\langle \boldsymbol{T}_\square, \text{vec}\left(\boldsymbol{v} \otimes \boldsymbol{v}\right) \text{vec}\left(\boldsymbol{v} \otimes \boldsymbol{v}\right)^\top \right\rangle \quad \text{subject to} \quad \|\boldsymbol{v}\| = 1$$

$\Updownarrow$ $\boldsymbol{X} = \text{vec}(\boldsymbol{v} \otimes \boldsymbol{v})\text{vec}(\boldsymbol{v} \otimes \boldsymbol{v})^\top$

$\Downarrow$ **induced symmetry** & $\|\boldsymbol{v}\| = 1$

$$\text{minimize} \quad \langle \boldsymbol{T}_\square, \boldsymbol{X} \rangle \quad \text{subject to} \quad \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \ \boldsymbol{X} \succeq \boldsymbol{0}, \text{rank}(\boldsymbol{X}) = 1$$

$\Updownarrow$ **convex relaxation**

$$\text{minimize} \quad \langle \boldsymbol{T}_\square, \boldsymbol{X} \rangle \quad \text{subject to} \quad \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \ \boldsymbol{X} \succeq \boldsymbol{0}$$

(*) The $p = 3$ tensor problem can be transformed into a $p = 4$ tensor problem [JMZ, '14].

SDP Relaxation

Linear constraints $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$

E.g. $\boldsymbol{v} = (v_1, v_2)^T$

$$\boldsymbol{X} = vec(\boldsymbol{v} \otimes \boldsymbol{v})\, vec(\boldsymbol{v} \otimes \boldsymbol{v})^T = \begin{bmatrix} v_1^2 \\ v_1 v_2 \\ v_2 v_1 \\ v_2^2 \end{bmatrix} \begin{bmatrix} v_1^2 & v_1 v_2 & v_2 v_1 & v_2^2 \end{bmatrix}$$

$$= \begin{bmatrix} v_1^4 & v_1^3 v_2 & v_2 v_1^3 & v_1^2 v_2^2 \\ v_1^3 v_2 & v_1^2 v_2^2 & v_1^2 v_2^2 & v_1 v_2^3 \\ v_2 v_1^3 & v_1^2 v_2^2 & v_1^2 v_2^2 & v_1 v_2^3 \\ v_1^2 v_2^2 & v_1 v_2^3 & v_1 v_2^3 & v_2^4 \end{bmatrix}$$

$\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$

- **spherical constr.** $1 = \|v\|_2^2 = v_1^2 + v_2^2$
  $= (v_1^2 + v_2^2)^2 = trace(\boldsymbol{X})$
- **hyper- symmetry** $\boldsymbol{X} = \begin{bmatrix} a & d & d & b \\ d & b & b & e \\ d & b & b & e \\ b & e & e & c \end{bmatrix}$

SDP Relaxation

$$\text{minimize} \quad \langle \boldsymbol{T}_{\square}, \boldsymbol{X} \rangle \quad \text{subject to} \quad \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \; \boldsymbol{X} \succeq \boldsymbol{0}$$

N.B.

- SDP relaxation proposed by [Jiang, Ma & Zhang, '14]
- Square reshaping trick for low-rank tensor recovery [MHW**G**, '14]

  $$\boldsymbol{T}_{i+(j-1)n,k+(l-1)n} \leftarrow \boldsymbol{\mathcal{T}}_{ijkl}$$

- Equivalent SDP (of reduced size) proposed by [Nie & Wang, '14] using moment-based convex relaxation
- Empirically, $\mathrm{rank}(\boldsymbol{X}^{\star}) = 1$ observed **almost always!**
  i.e. , SDP relaxation → solves nonconvex problem!

Solving SDP

Semidefinite programming (SDP)

$$\begin{aligned} \min \quad & \langle \boldsymbol{C}, \boldsymbol{X} \rangle \\ \text{s.t.} \quad & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b} \\ & \boldsymbol{X} \succeq \boldsymbol{0} \end{aligned}$$

Linear constraints:

$$\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$$

$$\Updownarrow \quad \mathcal{A}(\boldsymbol{X}) = \left(\langle \boldsymbol{A}_1, \boldsymbol{X} \rangle, \langle \boldsymbol{A}_2, \boldsymbol{X} \rangle \cdots, \langle \boldsymbol{A}_m, \boldsymbol{X} \rangle\right)^\top$$

$$\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i, \quad i \in [m]$$

$$\Updownarrow \quad \boldsymbol{A} = \Big(\mathrm{vec}\left(\boldsymbol{A}_1\right), \mathrm{vec}\left(\boldsymbol{A}_2\right), \cdots, \mathrm{vec}\left(\boldsymbol{A}_m\right)\Big)^\top$$

$$\boldsymbol{A}\mathrm{vec}(\boldsymbol{X}) = \boldsymbol{b}$$

Deriving the dual problem

Lagrangian function

$$\begin{aligned}\mathcal{L}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{S}) &= \langle \boldsymbol{C}, \boldsymbol{X} \rangle - \boldsymbol{y}^\top (\mathcal{A}(\boldsymbol{X}) - \boldsymbol{b}) - \langle \boldsymbol{X}, \boldsymbol{S} \rangle \\ &= \langle \boldsymbol{C} - \mathcal{A}^*(\boldsymbol{y}) - \boldsymbol{S}, \boldsymbol{X} \rangle + \boldsymbol{y}^\top \boldsymbol{b}\end{aligned}$$

$$\mathcal{A}(\boldsymbol{X}) = \Big(\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \Big)_{i \in [m]}$$

$$\mathcal{A}^*(\boldsymbol{y}) = \sum_{i \in [m]} y_i \boldsymbol{A}_i$$

Dual problem

$$\max_{\boldsymbol{y},\, \boldsymbol{S} \succeq \boldsymbol{0}} \ \min_{\boldsymbol{X}} \quad \mathcal{L}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{S}) = \langle \boldsymbol{C} - \mathcal{A}^*(\boldsymbol{y}) - \boldsymbol{S}, \boldsymbol{X} \rangle + \boldsymbol{y}^\top \boldsymbol{b}$$

$$\begin{aligned}\max_{\boldsymbol{y},\, \boldsymbol{S}} \quad & \boldsymbol{b}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} = \boldsymbol{C} \\ & \boldsymbol{S} \succeq \boldsymbol{0}\end{aligned}$$

$$\begin{aligned}\min_{\boldsymbol{y},\, \boldsymbol{S}} \quad & -\boldsymbol{b}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} = \boldsymbol{C} \\ & \boldsymbol{S} \succeq \boldsymbol{0}\end{aligned}$$

Augmented Lagrangian Method

$$(D) \quad \min_{\boldsymbol{y},\, \boldsymbol{S}} \; -\boldsymbol{b}^\top \boldsymbol{y} \quad \text{s.t.} \quad \mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} = \boldsymbol{C}, \; \boldsymbol{S} \succeq \boldsymbol{0}$$

Augmented Lagrangian function

$$\mathcal{L}_\mu(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{S}) = -\boldsymbol{y}^\top \boldsymbol{b} + \langle \mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} - \boldsymbol{C}, \boldsymbol{X} \rangle + \frac{1}{2\mu} \|\mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} - \boldsymbol{C}\|_F^2$$

Augmented Lagrangian method (ALM)

***k*-th iteration:**

compute $\quad (\boldsymbol{y}^{k+1}, \boldsymbol{S}^{k+1}) \in \arg\min_{\boldsymbol{y}, \boldsymbol{S} \succeq \boldsymbol{0}} \; \mathcal{L}_\mu(\boldsymbol{X}^k, \boldsymbol{y}, \boldsymbol{S})$

update $\quad \boldsymbol{X}^{k+1} = \boldsymbol{X}^k + \left(\mathcal{A}^*(\boldsymbol{y}^{k+1}) + \boldsymbol{S}^{k+1} - \boldsymbol{C} \right) \Big/ \mu$

- minimizing $\mathcal{L}_\mu(\boldsymbol{X}^k, \boldsymbol{y}, \boldsymbol{S})$ jointly over $(\boldsymbol{y}, \boldsymbol{S})$ is hard!

Alternating Direction Method of Multipliers (ADMM)

Remedy: alternating direction

minimize $\mathcal{L}_\mu(\boldsymbol{X}^k, \boldsymbol{y}, \boldsymbol{S})$ along ***y***-direction and ***S***-direction alternatively

ADMM

_k_-th iteration:

y-update: $\boldsymbol{y}^{k+1} \leftarrow \arg\min_{\boldsymbol{y}} \ \mathcal{L}_\mu(\boldsymbol{X}^k, \boldsymbol{y}, \boldsymbol{S}^k)$

S-update: $\boldsymbol{S}^{k+1} \leftarrow \arg\min_{\boldsymbol{S}\succeq \boldsymbol{0}} \ \mathcal{L}_\mu(\boldsymbol{X}^k, \boldsymbol{y}^{k+1}, \boldsymbol{S})$

X-update: $\boldsymbol{X}^{k+1} = \boldsymbol{X}^k + \Big(\mathcal{A}^*(\boldsymbol{y}^{k+1}) + \boldsymbol{S}^{k+1} - \boldsymbol{C}\Big)\Big/\mu$

Each step is (relatively) easy to compute!

Update $\boldsymbol{y}$

$$\boldsymbol{y}^{k+1} \leftarrow \arg\min_{\boldsymbol{y}} \quad \mathcal{L}_\mu(\boldsymbol{X}^k, \boldsymbol{y}, \boldsymbol{S}^k)$$

$$\mathcal{L}_\mu(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{S}) = -\boldsymbol{y}^\top \boldsymbol{b} + \langle \mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} - \boldsymbol{C}, \boldsymbol{X} \rangle + \frac{1}{2\mu} \|\mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} - \boldsymbol{C}\|_F^2$$

First-order optimality condition:

$$\nabla_{\boldsymbol{y}} \mathcal{L}_\mu = -\boldsymbol{b} + \mathcal{A}(\boldsymbol{X}^k) + \frac{1}{\mu} \mathcal{A}\left(\mathcal{A}^*(\boldsymbol{y}^{k+1}) + \boldsymbol{S}^k - \boldsymbol{C}\right)$$

$\Updownarrow$ assume $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ is onto (surjective)

$$\boldsymbol{y}^{k+1} = -(\mathcal{A}\mathcal{A}^*)^{-1} \left(\mu \left(\mathcal{A}(\boldsymbol{x}^k - b)\right) + \mathcal{A}\left(\boldsymbol{S}^k - \boldsymbol{C}\right)\right)$$

Update $\boldsymbol{S}$

$$\mathcal{L}_\mu(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{S}) = -\boldsymbol{y}^\top \boldsymbol{b} + \langle \mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} - \boldsymbol{C}, \boldsymbol{X} \rangle + \frac{1}{2\mu} \|\mathcal{A}^*(\boldsymbol{y}) + \boldsymbol{S} - \boldsymbol{C}\|_F^2$$

$$\begin{aligned} \boldsymbol{S}^{k+1} &= \arg\min_{\boldsymbol{S} \succeq \boldsymbol{0}} \; \mathcal{L}_\mu(\boldsymbol{X}^k, \boldsymbol{y}^{k+1}, \boldsymbol{S}) \\ &= \arg\min_{\boldsymbol{S} \succeq \boldsymbol{0}} \; \left\|\boldsymbol{S} - \boldsymbol{V}^{k+1}\right\|_F^2 \\ &= \boldsymbol{Q}_+ \boldsymbol{\Sigma}_+ \boldsymbol{Q}_+^\top \end{aligned}$$

with

$$\begin{aligned} \boldsymbol{V}^{k+1} &= \boldsymbol{C} - \mathcal{A}^*(\boldsymbol{y}^{k+1}) - \mu \boldsymbol{X}^k \\ &= \begin{bmatrix} \boldsymbol{Q}_+ & \boldsymbol{Q}_- \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_+ & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_- \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_+^\top \\ \boldsymbol{Q}_-^\top \end{bmatrix} \end{aligned}$$

(Eigen-Decomp.)

Update muliplier *X*

$$
\begin{aligned}
\boldsymbol{X}^{k+1} &= \boldsymbol{X}^{k} + \left(\mathcal{A}^*(\boldsymbol{y}^{k+1}) + \boldsymbol{S}^{k+1} - \boldsymbol{C}\right)\Big/\mu \\
&= \frac{1}{\mu}\left(\boldsymbol{S}^{k+1} - (\boldsymbol{C} - \mathcal{A}\left(\boldsymbol{y}^{k+1}\right) - \mu\boldsymbol{X}^{k})\right) \\
&= \frac{1}{\mu}(\boldsymbol{S}^{k+1} - \boldsymbol{V}^{k+1}) \\
&= -\frac{1}{\mu}\boldsymbol{Q}_{-}\boldsymbol{\Sigma}_{-}\boldsymbol{Q}_{-}^{\top}
\end{aligned}
$$

Convergence

Semidefinite programming (SDP)

$$\text{minimize} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle \quad \text{subject to} \quad \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \quad \boldsymbol{X} \succeq \boldsymbol{0}$$

Assumption:

- $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ is onto
- $\exists \; \boldsymbol{X}_0$ s.t. $\mathcal{A}(\boldsymbol{X}_0) = \boldsymbol{b}$ and $\boldsymbol{X}_0 \succ \boldsymbol{0}$ (Slater's condition)

Convergence result

Theorem (WGY, '10)

From any starting point,

$$\{(\boldsymbol{X}^k, \boldsymbol{y}^k, \boldsymbol{S}^k)\} \to \text{a primal and dual solution } \{(\boldsymbol{X}^\star, \boldsymbol{y}^\star, \boldsymbol{S}^\star)\}$$

Another ADMM for Tensor Problem

Tensor robust principal component anaysis (T-RPCA)

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$$

Structural assumptions:

$\boldsymbol{\mathcal{L}}$: low rank in each mode ($\boldsymbol{\mathcal{L}}_{(\boldsymbol{k})}$)

i.e. rank($\boldsymbol{\mathcal{L}}_{(\boldsymbol{k})}$) is small for all $k \in [K]$

$\boldsymbol{\mathcal{S}}$: sparsely supported

i.e. cardinality{$\boldsymbol{\mathcal{S}}$: $\boldsymbol{\mathcal{S}} \neq 0$} is small

Problem: Given $\boldsymbol{\mathcal{X}}$, recover $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{S}}$.

T-RPCA

Convex surrogates

$$\mathrm{rank}(\boldsymbol{\mathcal{L}}_{(\boldsymbol{i})}) \Longrightarrow \left\|\boldsymbol{\mathcal{L}}_{(\boldsymbol{i})}\right\|_* := \sum \boldsymbol{\sigma}_{\boldsymbol{j}}(\boldsymbol{\mathcal{L}}_{(\boldsymbol{i})})$$

$$\mathrm{cardinality}(\boldsymbol{\mathcal{S}}) \Longrightarrow \|\boldsymbol{\mathcal{S}}\|_1 := \sum |\boldsymbol{\mathcal{S}}_{\boldsymbol{i_1 i_2 \cdots i_K}}|$$

Convex optimization

$$\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{S}}} \quad \sum \lambda_i \left\|\boldsymbol{\mathcal{L}}_{(i)}\right\|_* + \|\boldsymbol{\mathcal{S}}\|_1$$

$$\text{s.t.} \quad \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{X}}$$

N.B.

- generalizes matrix robust PCA [CLMW, 11]
- theoretical guarantees [MHWG, '14] [HMGW, '15]

T-RPCA

Variable Splitting

$$\sum_i \lambda_{\boldsymbol{i}} \left\| \boldsymbol{\mathcal{L}}_{(\boldsymbol{i})} \right\|_*$$

$$\Downarrow$$

$$\boldsymbol{\mathcal{L}}$$

$$\boldsymbol{\mathcal{L}}_1 = \boldsymbol{\mathcal{L}}_2 = \cdots = \boldsymbol{\mathcal{L}}_K$$

$$\sum_i \lambda_i \left\| \boldsymbol{\mathcal{L}}_{i,(i)} \right\|_*$$

T-RPCA

Reformulated problem

$$\min_{\{\boldsymbol{L}_i\},\boldsymbol{S}} \quad \sum \lambda_i \left\|\boldsymbol{\mathcal{L}}_{i,(i)}\right\|_* + \|\boldsymbol{\mathcal{S}}\|_1$$

$$\text{s.t.} \quad \boldsymbol{\mathcal{L}}_i + \boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{X}}, \ \forall\, i \in [K]$$

Augmented Lagrangian function $\mathcal{L}_\mu\Big(\{\boldsymbol{\mathcal{L}}_i\}, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\Lambda}_i\} \Big)$

$$\sum_{i\in[K]} \lambda_i \left\|\boldsymbol{\mathcal{L}}_{i,(i)}\right\|_* + \|\boldsymbol{\mathcal{S}}\|_1 + \sum_{i\in[K]} \Big(-\langle \boldsymbol{\Lambda}_i, \boldsymbol{\mathcal{L}}_i + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}} \rangle + \frac{1}{2\mu} \|\boldsymbol{\mathcal{L}}_i + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}\|_F^2 \Big)$$

ADMM *k*-th iteration:

$\boldsymbol{\mathcal{L}}_i$ -update: $\left\{\boldsymbol{\mathcal{L}}_i^{k+1}\right\} \leftarrow \underset{\{\boldsymbol{\mathcal{L}}_i\}}{\arg\min} \ \mathcal{L}_\mu\Big(\{\boldsymbol{\mathcal{L}}_i\}, \boldsymbol{\mathcal{S}}^k, \{\boldsymbol{\Lambda}_i^k\} \Big)$

$\boldsymbol{\mathcal{S}}$-update: $\boldsymbol{\mathcal{S}}^{k+1} \leftarrow \underset{\boldsymbol{\mathcal{E}}}{\arg\min} \ \mathcal{L}_\mu\Big(\left\{\boldsymbol{\mathcal{L}}_i^{k+1}\right\}, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\Lambda}_i^k\} \Big)$

$\boldsymbol{\Lambda}_i$-update: $\boldsymbol{\Lambda}_i^{k+1} \leftarrow \boldsymbol{\Lambda}_i^k - \frac{1}{\mu}(\boldsymbol{\mathcal{L}}_i^{k+1} + \boldsymbol{\mathcal{S}}^{k+1} - \boldsymbol{\mathcal{X}}) \quad \forall i \in [K]$

References

[ZG] Zhang, T. and Golub G. H.. "Rank-one approximation to high order tensors." *SIAM Journal on Matrix Analysis and Applications* 23.2 (2001): 534-550.

[AGHKT] Anandkumar, A., et al. "Tensor decompositions for learning latent variable models." *The Journal of Machine Learning Research* 15.1 (2014): 2773-2832.

[HS] Hsu, D. and Sham M. K.. "Learning mixtures of spherical gaussians: moment methods and spectral decompositions." *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. ACM, 2013.

[TB] Kolda, T. G. and Bader B. W. Bader.. "Tensor decompositions and applications." *SIAM review* 51.3 (2009): 455-500.

[JMZ] Jiang, B., Ma, S. and Zhang S.. "Tensor principal component analysis via convex optimization." *Mathematical Programming*150.2 (2015): 423-457.

[NW] Nie, J. and Wang, L.. "Semidefinite relaxations for best rank-1 tensor approximations." *SIAM Journal on Matrix Analysis and Applications* 35.3 (2014): 1155-1179.

[CLMW] Candès, E. J., et al. "Robust principal component analysis?."*Journal of the ACM (JACM)* 58.3 (2011): 11.

References

[WGY] Wen, Z., Goldfarb, D., & Yin, W. (2010). Alternating direction augmented Lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, *2*(3-4), 203-230.

[MHWG] Mu, C., et al. "Square deal: Lower bounds and improved relaxations for tensor recovery." *Proceedings of the international conference on Machine learning*. ACM, 2014

[GQ] Goldfarb, D., & Qin, Z.. "Robust low-rank tensor recovery: Models and algorithms." *SIAM Journal on Matrix Analysis and Applications* 35.1 (2014): 225-253.

[HMGW] Huang, B., et al. "Provable models for robust low-rank tensor completion." *Pacific Journal of Optimization* 11 (2015): 339-364.

[MHG] Mu, C., Hsu, D., & Goldfarb, D.. "Successive Rank-One Approximations for Nearly Orthogonally Decomposable Symmetric Tensors." *SIAM Journal on Matrix Analysis and Applications* 36.4 (2015): 1638-1659.