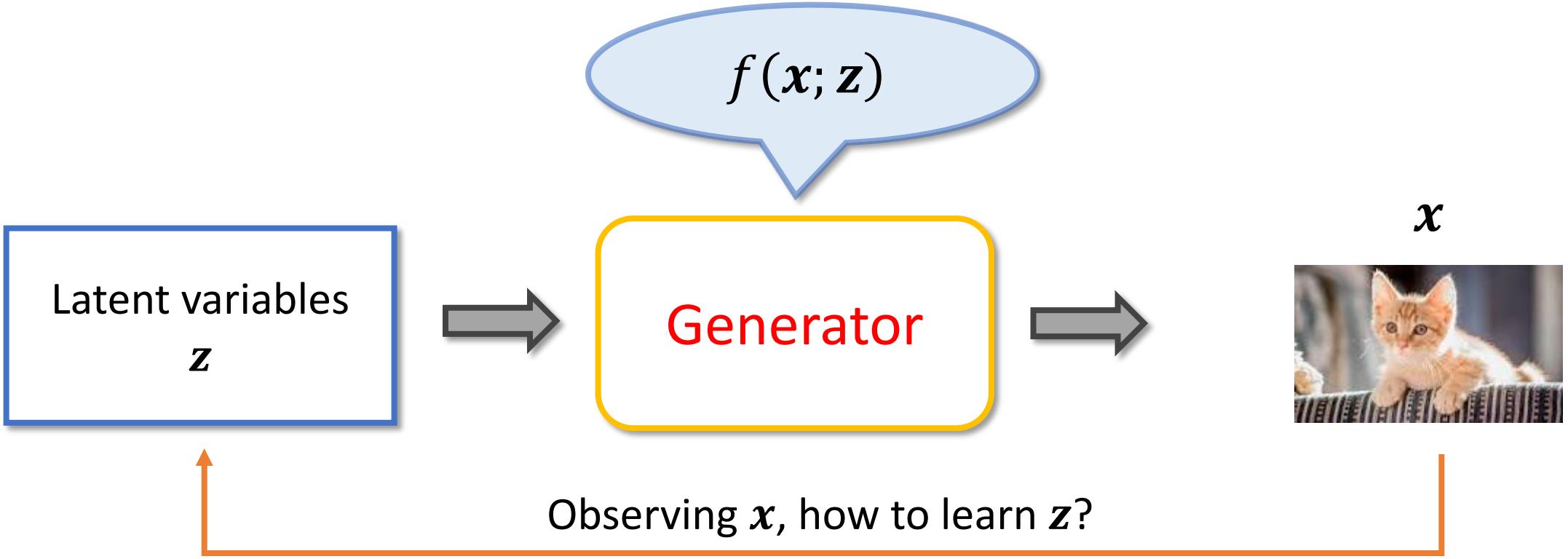
CS 57800 – Statistical Machine Learning

Lecture 2.5 Latent Variables Learning (II)

https://ruizhezhang.com/course_fall_2026.html

Latent Variable Models



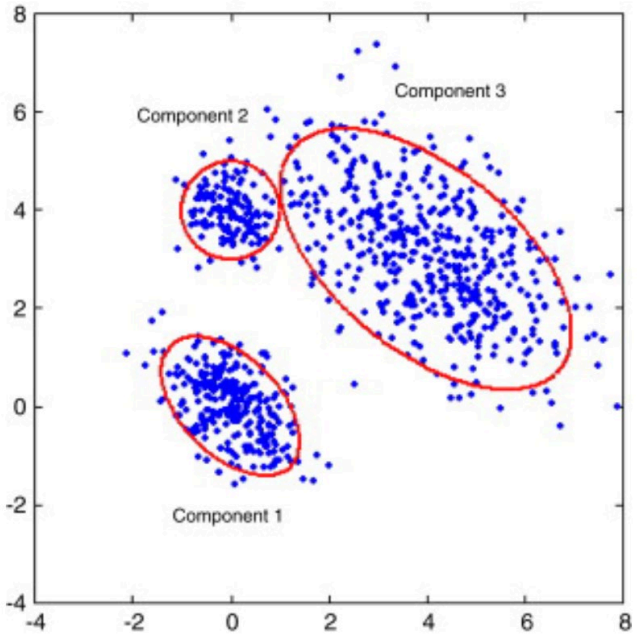
Outline

- Matrix Completion

- Method of Moments

Two Canonical Examples

- **Mixture of Gaussians**
  - Each point is generated by an (unknown) cluster



- **Topic models**
  - "bag of words" model for documents
  - documents have one (or more) topics

What is the statistical efficiency of the estimator we find?

Heuristics: k-means, expectation–maximization (EM), …

Two Canonical Examples

(spherical) Mixture of Gaussians

- k means: $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$
- Sample cluster $i \in [k]$ with probability w_i
- Observe $\boldsymbol{x}$, with spherical noise,

$$\boldsymbol{x} = \boldsymbol{\mu}_i + \eta, \qquad \eta \sim \mathcal{N}\left(0, \sigma_i^2 I\right)$$

- Dataset: multiple points

(single) Topic Model

- k topics: $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$
- Sample topic $i \in [k]$ with probability w_i
- Observe m (exchangeable) words

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

- Dataset: multiple m-word documents

How to learn the parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k,\ w_1, \ldots, w_k$ (and σ_i^2)?

Detour: Topic Model

A concrete example



Documents with explicit topics
e.g., newspaper, magazine, …

bag-of-words

$w_1 = 0.6$ $w_2 = 0.4$

Word	Topic 1: "sports"	Topic 2: "science"
team	0.5	0.05
game	0.35	0.1
experiment	0.1	0.5
data	0.05	0.35

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0.5 \\ 0.35 \\ 0.1 \\ 0.05 \end{bmatrix} \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} 0.05 \\ 0.1 \\ 0.5 \\ 0.35 \end{bmatrix}$$

Detour: Topic Model

A concrete example



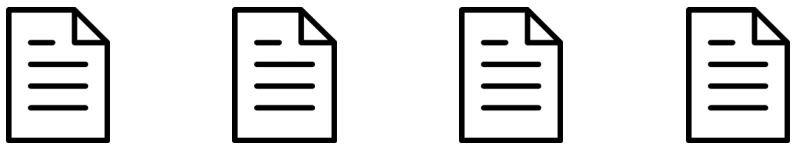
Documents with explicit topics
e.g., newspaper, magazine, …

bag-of-words

	$w_1 = 0.6$	$w_2 = 0.4$
Word	**Topic 1: "sports"**	**Topic 2: "science"**
team	0.5	0.05
game	0.35	0.1
experiment	0.1	0.5
data	0.05	0.35

Generation:

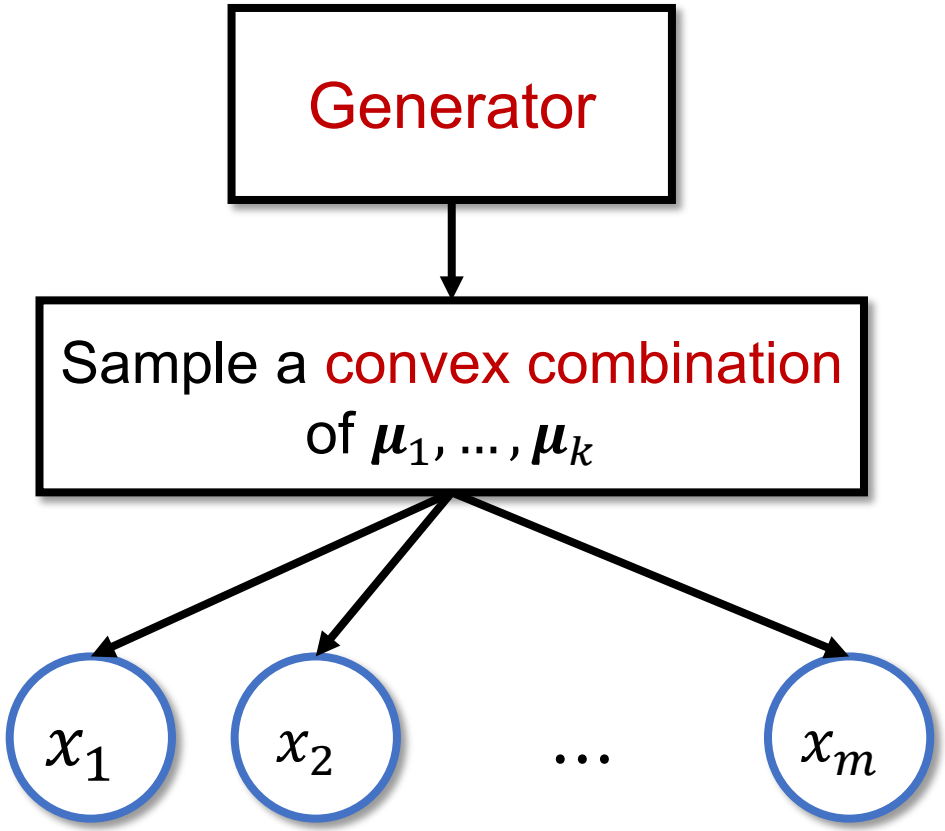
$$\Pr[x_1 = \text{"game"}, x_2 = \text{"data"} | \text{Topic 1}] = 0.35 \cdot 0.05$$

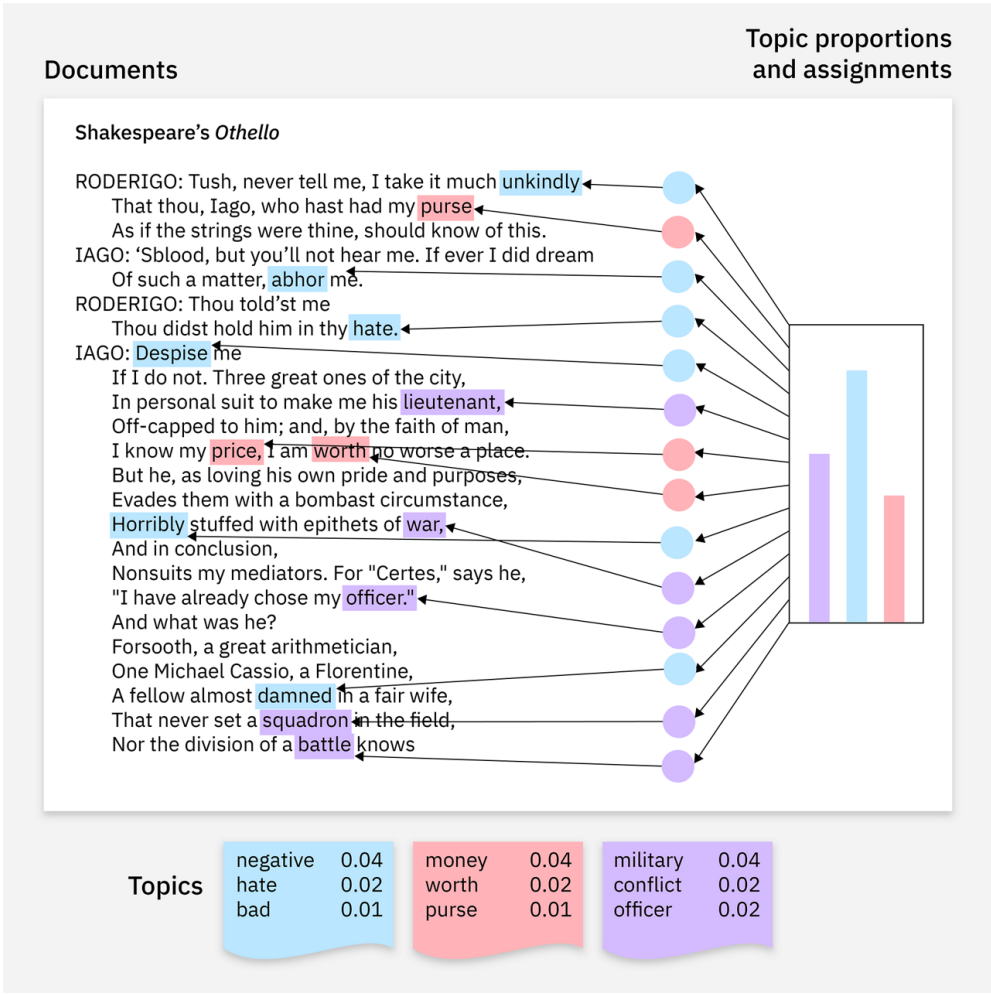


$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0.5 \\ 0.35 \\ 0.1 \\ 0.05 \end{bmatrix} \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} 0.05 \\ 0.1 \\ 0.5 \\ 0.35 \end{bmatrix}$$

Detour: Topic Model

Real-world documents are mixtures of topics





Two Canonical Examples

(spherical) Mixture of Gaussians

- k means: $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$
- Sample cluster $i \in [k]$ with probability w_i
- Observe $\boldsymbol{x}$, with spherical noise,

$$\boldsymbol{x} = \boldsymbol{\mu}_i + \boldsymbol{\eta}, \qquad \boldsymbol{\eta} \sim \mathcal{N}\left(0, \sigma_i^2 I\right)$$

- Dataset: multiple points

(single) Topic Model

- k topics: $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$
- Sample topic $i \in [k]$ with probability w_i
- Observe m (exchangeable) words

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

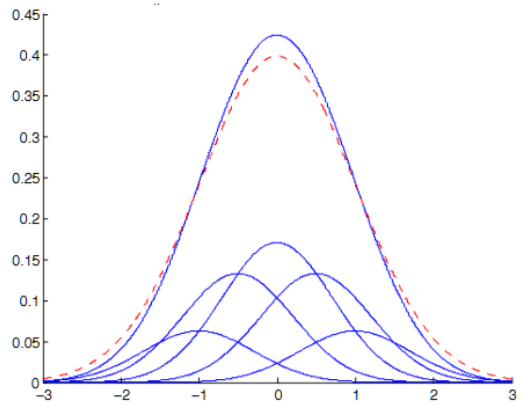
- Dataset: multiple m-word documents

How to learn the parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k,\ w_1, \ldots, w_k$ (and σ_i^2)?

Limits of Learning

Information-theoretical lower bound:

- **Moitra-Valiant '11:** need exponential in k many samples to learn a GMM with k components, even in 1-d



Fun fact: A discrete Gaussian mixture is almost Gaussian

Let $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Then

$$\frac{1}{Z} \sum_{i \in \mathbb{Z}} \phi(i/k) \cdot \mathcal{N}(i/k, 1) \approx_{e^{-\Omega(k^2)}} \mathcal{N}(0,2)$$

Computational lower bound:

- **Arora-Ge-Moitra '12:** computing the maximum likelihood estimation for learning the parameter of a topic model, more specifically, a Latent Dirichlet Allocation (LDA) model, is **NP**-hard

Limits of Learning

Information-theoretical lower bound:

- **Moitra-Valiant '11:** need exponential in k many samples to learn a GMM with k components, even in 1-d

Computational lower bound:

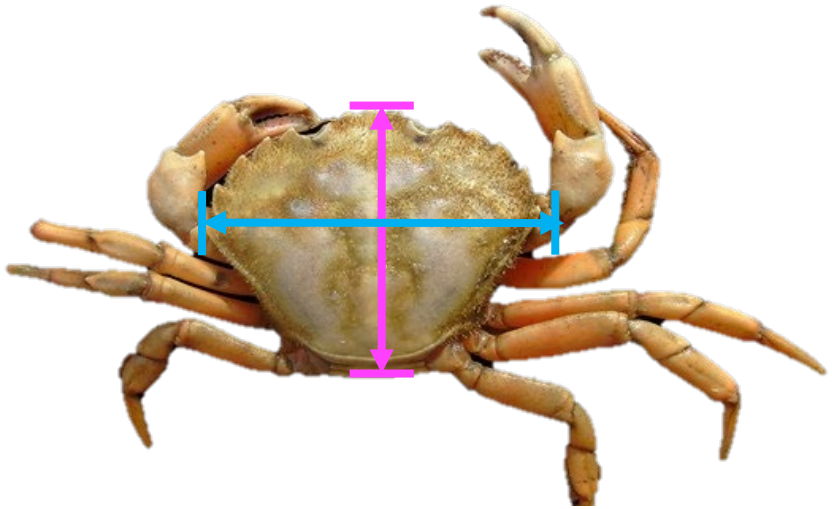
- **Arora-Ge-Moitra '12:** computing the maximum likelihood estimation for learning the parameter of a topic model, more specifically, a Latent Dirichlet Allocation (LDA) model, is **NP**-hard

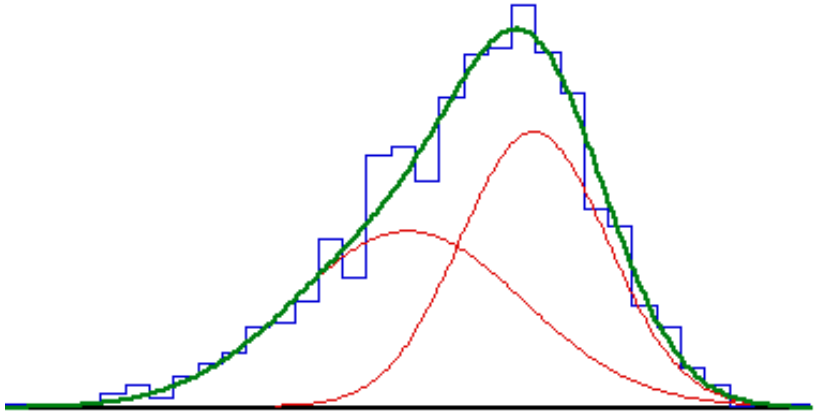
Are there statistically and computationally efficient methods?

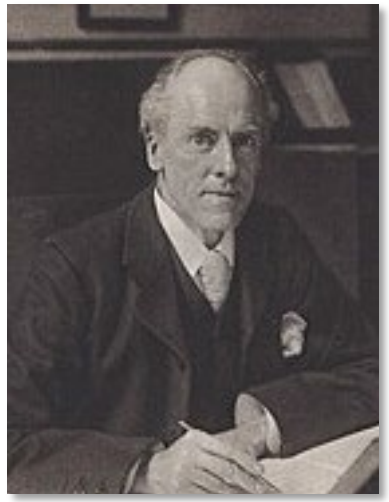
- Under what assumptions and models?
- How general?

Method of moments

- A simple approach based on linear algebra
  - For non-convex problems
- Extensions to richer settings
  - LDA, Hidden Markov model (HMM), etc.
- In high dimensions, GMMs and topic models can be efficiently learnable
- Pearson's crab data (1894)







Karl Pearson
(1857-1936)

The Method of Moments

- GMM moments:

$$\mathbb{E}[\boldsymbol{x}], \qquad \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}], \qquad \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}], \qquad \ldots$$

- Topic model moments:

$$\Pr[x_1], \qquad \Pr[x_1, x_2], \qquad \Pr[x_1, x_2, x_3], \qquad \ldots$$

- These moment data can be estimated from observed data
- **Identifiability problem:** with exact moments, which order suffices?
  - Can we design efficient algorithms?

Kronecker product

For $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{p \times q}$,

$$(A \otimes B)_{ik,j\ell} = A_{ij} B_{k\ell}$$

The Method of Moments

Setups

- There are k clusters/topics and d dimensions/words. $d \gg k$
- For GMMs,
  - $\boldsymbol{x} = \boldsymbol{\mu}_i + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \sigma_i^2 I)$
  - $\mathbb{E}[\boldsymbol{x} | \text{cluster } i] = \boldsymbol{\mu}_i$
- For topic models,
  - Use one-hot encoding for words: $\boldsymbol{x}_1 = [0 \quad 1 \quad 0 \quad \cdots]^\mathsf{T}$
  - The $\boldsymbol{\mu}_i$'s are probability vectors
  - $\Pr[x_1 | \text{topic } i] = \mathbb{E}[\boldsymbol{x}_1 | \text{topic } i] = \boldsymbol{\mu}_i$

The Method of Moments

First moment

GMM

$$\mathbb{E}[\boldsymbol{x}] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i$$

Topic model

$$\Pr[x_1] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i$$

Not identifiable: d nums, $dk + k$ vars

The Method of Moments

Second moment

GMM

$$\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] = \mathbb{E}[(\boldsymbol{\mu}_i + \boldsymbol{\eta}) \otimes (\boldsymbol{\mu}_i + \boldsymbol{\eta})]$$
$$= \mathbb{E}[\boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i] + \mathbb{E}[\boldsymbol{\eta} \otimes \boldsymbol{\eta}]$$
$$= \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i + \sigma^2 I$$

- A full-rank matrix

Topic model

By exchangeability,

$$\Pr[x_1, x_2] = \mathbb{E}\big[\mathbb{E}[\boldsymbol{x}_1|\text{topic}] \otimes \mathbb{E}[\boldsymbol{x}_2|\text{topic}]\big]$$
$$= \sum_{i=1}^{k} \boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1$$

- A low-rank matrix!

Still not identifiable

The Method of Moments

Second moment

Topic model

$$\Pr[x_1, x_2] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

Model A: $w_1 = 1/3\,, \boldsymbol{\mu}_1 = \begin{bmatrix} 1/6 \\ 5/6 \end{bmatrix}$ and $w_2 = 2/3\,, \boldsymbol{\mu}_2 = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$

Model B: $w_1 = 2/3\,, \boldsymbol{\mu}_1 = \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}$ and $w_2 = 1/3\,, \boldsymbol{\mu}_2 = \begin{bmatrix} 5/6 \\ 1/6 \end{bmatrix}$

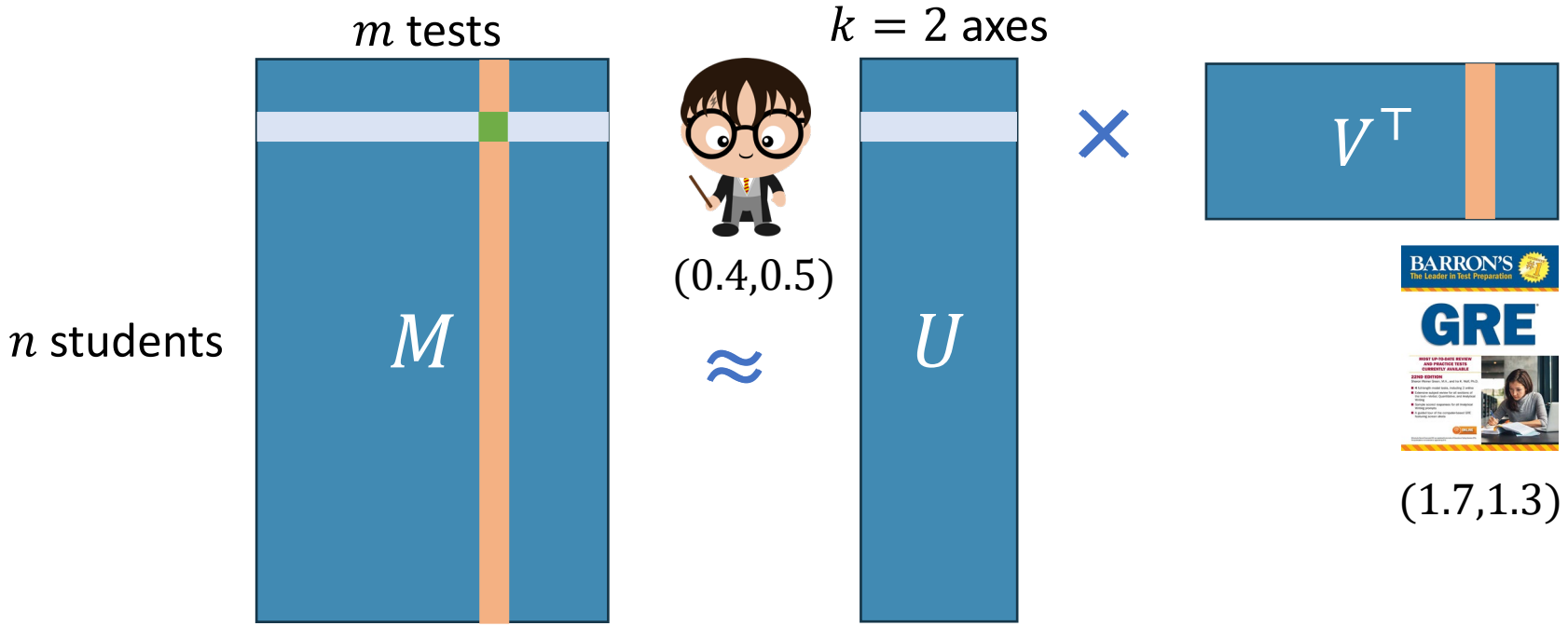
$$\Pr[x_1] = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

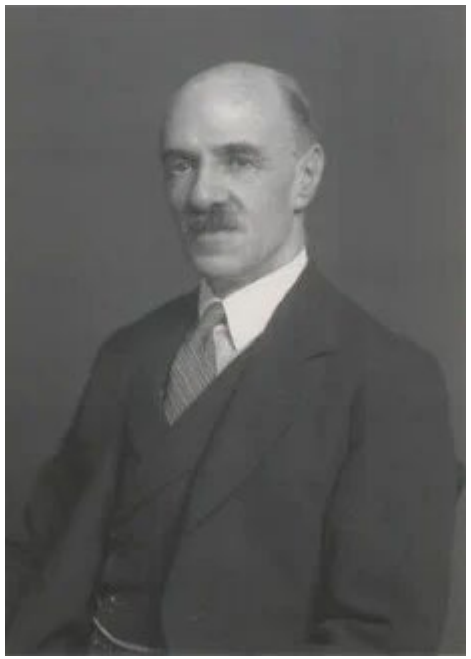
$$\Pr[x_1, x_2] = \begin{bmatrix} 11/36 & 7/36 \\ 7/36 & 11/36 \end{bmatrix}$$

Detour: the Rotation Problem

Factor analysis is a statistical method, pioneered by Charles Spearman, that explains observed correlations among many variables by modeling them as combinations of a few underlying latent factors

Suppose a psychologist has the hypothesis that there are two kinds of intelligence, "verbal" and "mathematical".



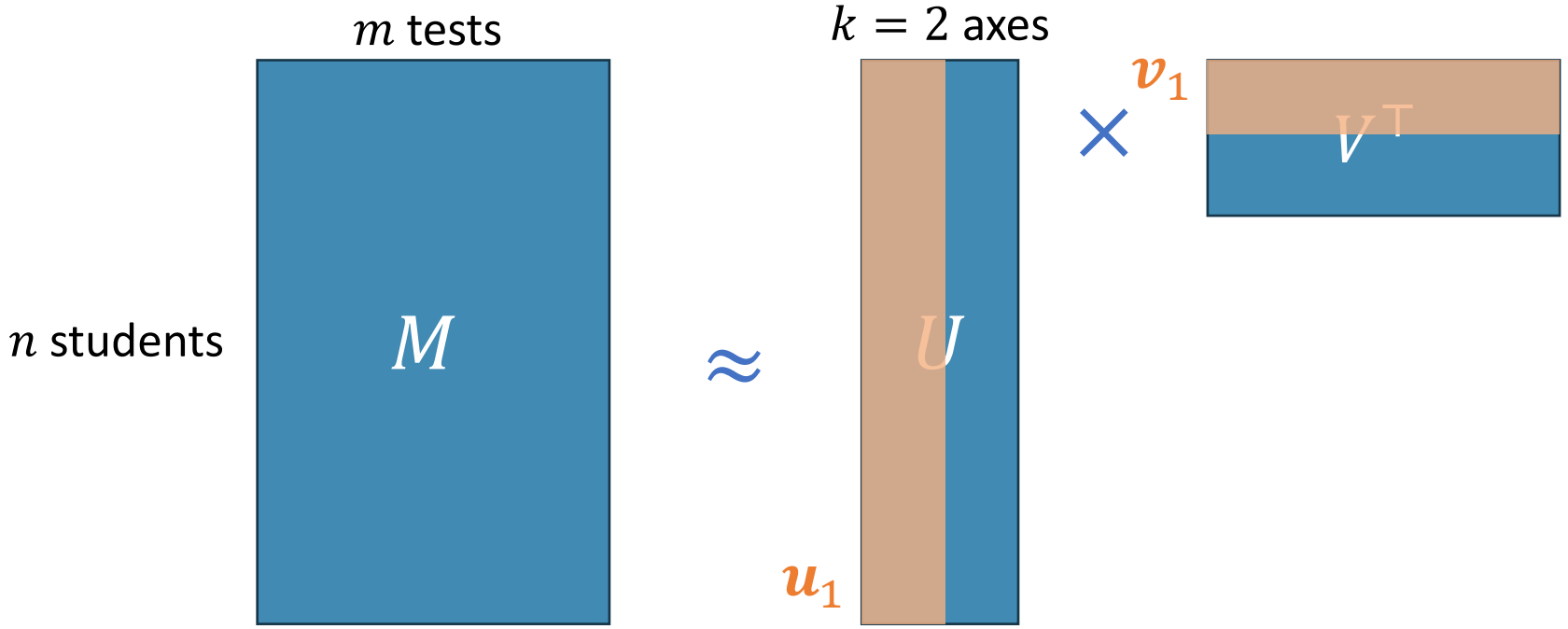


Charles Spearman
(1863-1945)

Detour: the Rotation Problem

Factor analysis is a statistical method, pioneered by Charles Spearman, that explains observed correlations among many variables by modeling them as combinations of a few underlying latent factors.

Suppose a psychologist has the hypothesis that there are two kinds of intelligence, "verbal" and "mathematical".



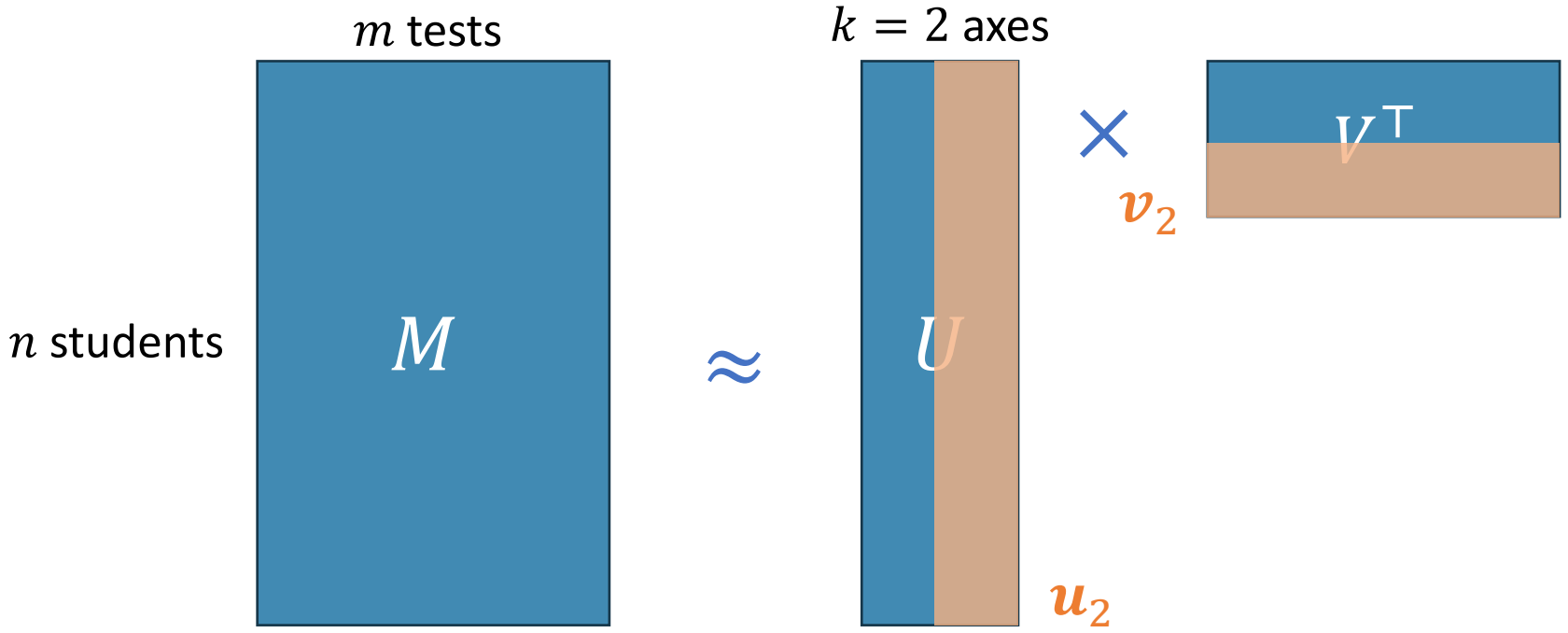


Charles Spearman
(1863-1945)

Detour: the Rotation Problem

Factor analysis is a statistical method, pioneered by Charles Spearman, that explains observed correlations among many variables by modeling them as combinations of a few underlying latent factors.

Suppose a psychologist has the hypothesis that there are two kinds of intelligence, "verbal" and "mathematical".





Charles Spearman
(1863-1945)

Detour: the Rotation Problem

Factor analysis is a statistical method, pioneered by Charles Spearman, that explains observed correlations among many variables by modeling them as combinations of a few underlying latent factors.

Suppose a psychologist has the hypothesis that there are two kinds of intelligence, "verbal" and "mathematical".

$$M = UV^\top = \sum_{i \in [k]} \boldsymbol{u}_i \boldsymbol{v}_i^\top$$



Charles Spearman
(1863-1945)

The rotation problem: this factorization is not unique

- Let $\tilde{U} \leftarrow UO, \tilde{V} \leftarrow VO$, where $O \in \mathbb{R}^{k \times k}$ is any orthogonal matrix
- $\tilde{U}\tilde{V}^\top = UOO^\top V^\top = UV^\top$

Detour: the Rotation Problem

Factor analysis is a statistical method, pioneered by Charles Spearman, that explains observed correlations among many variables by modeling them as combinations of a few underlying latent factors.

Suppose a psychologist has the hypothesis that there are two kinds of intelligence, "verbal" and "mathematical".

$$M = UV^\top = \sum_{i \in [k]} \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

The rotation problem: this factorization is not unique

- Unless we put some additional assumptions, such as $\mathrm{rank}(M) = 1$, $\{u_k\}$ and $\{v_k\}$ are orthogonal, $\{u_k, v_k\}$ only have non-negative entries…



Charles Spearman
(1863-1945)

The Method of Moments

Third moment ($d \times d \times d$ tensor)

Topic model

$$\Pr[x_1, x_2] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$\Pr[x_1, x_2, x_3] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$O(d^3)$ nums, $dk + k$ vars. Possibly identifiable?

The Method of Moments

Third moment ($d \times d \times d$ tensor)

Topic model

$$\Pr[x_1, x_2] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$\Pr[x_1, x_2, x_3] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

Model A: $w_1 = 1/3\,, \boldsymbol{\mu}_1 = \begin{bmatrix} 1/6 \\ 5/6 \end{bmatrix}$ and $w_2 = 2/3\,, \boldsymbol{\mu}_2 = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$ $\quad \Pr_A[x_1 = x_2 = x_3 = 1] = 43/216$

Model B: $w_1 = 2/3\,, \boldsymbol{\mu}_1 = \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}$ and $w_2 = 1/3\,, \boldsymbol{\mu}_2 = \begin{bmatrix} 5/6 \\ 1/6 \end{bmatrix}$ $\quad \Pr_B[x_1 = x_2 = x_3 = 1] = 47/216$

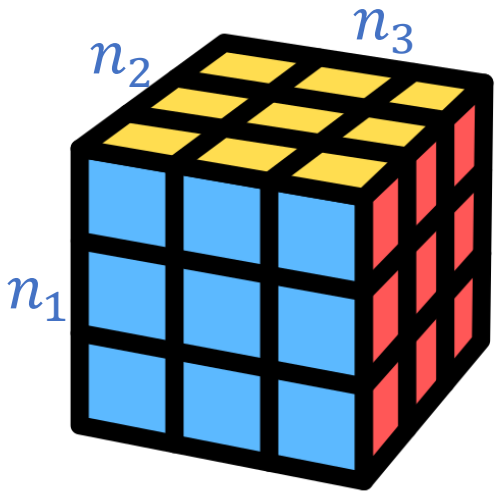
Detour: Tensor and Linear Algebra

- A rank one, third-order tensor T is the tensor product of three vectors $\boldsymbol{u} \in \mathbb{R}^{n_1}$, $\boldsymbol{v} \in \mathbb{R}^{n_2}$ and $\boldsymbol{w} \in \mathbb{R}^{n_3}$, and its entries are

$$T_{ijk} = u_i v_j w_k, \qquad i \in [n_1], j \in [n_2], k \in [n_3]$$

We denote it as

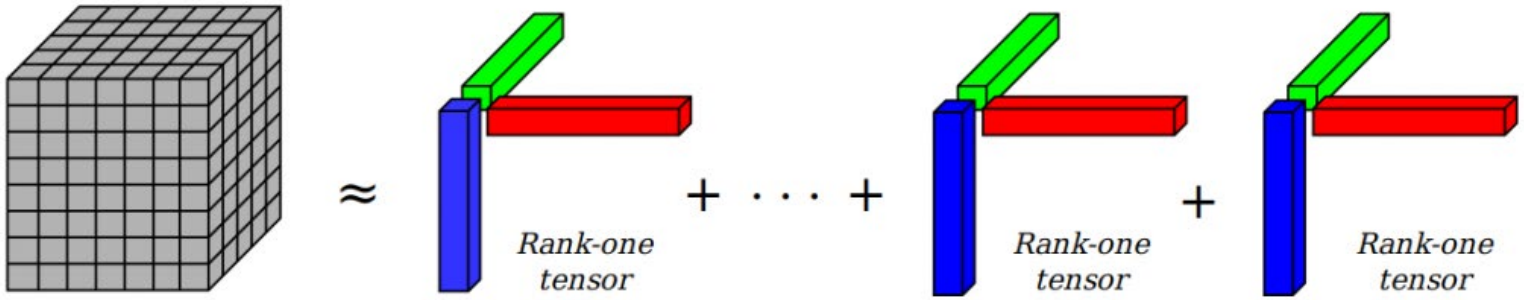
$$T = \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}$$



- The rank of a third-order tensor T is the smallest integer r so that we can write

$$T = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)}$$

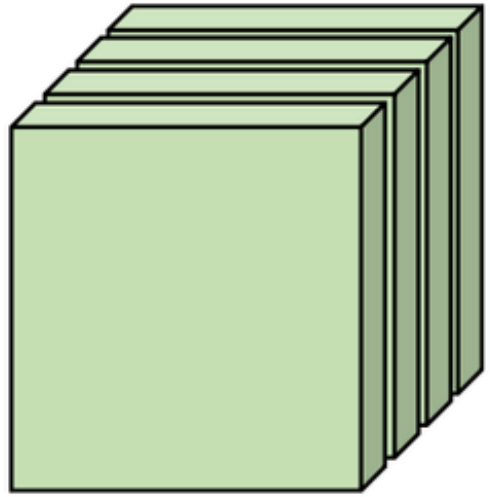
This decomposition is often called a CP decomposition or CANDECOMP/PARFAC decomposition



Detour: Tensor and Linear Algebra

We can view tensor T as a stacked collection of n_1-by-n_2 matrices:

$$T(:,:,1), T(:,:,2), \ldots, T(:,:,n_3)$$



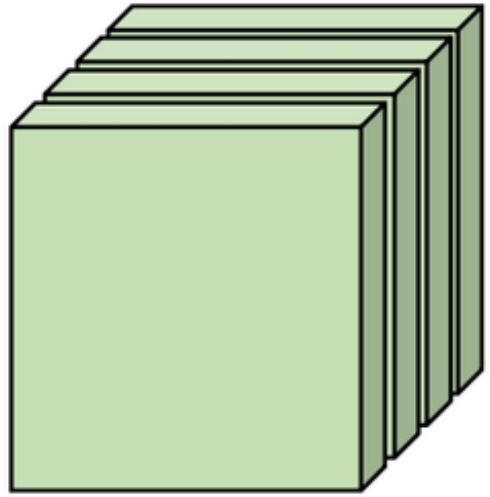
Claim 1. If $\text{rank}(T) \leq r$, then for all $a \in [n_3]$, $\text{rank}\big(T(:,:,a)\big) \leq r$.

Proof.

$$T = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)} \quad \Longrightarrow T(:,:,a) = \sum_{i=1}^{r} w_a^{(i)} \underbrace{\big(\boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)}\big)}_{\text{rank-1 matrix}}$$

■

Detour: Tensor and Linear Algebra



We can view tensor T as a stacked collection of n_1-by-n_2 matrices:

$$T(:,:,1), T(:,:,2), \ldots, T(:,:,n_3)$$

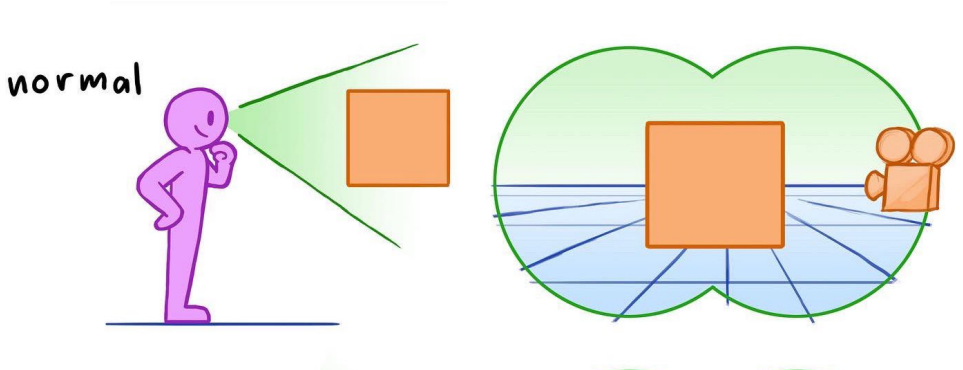
Claim 1. If $\text{rank}(T) \leq r$, then for all $a \in [n_3]$, $\text{rank}\big(T(:,:,a)\big) \leq r$.

Claim 2. Consider a tensor $T = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)}$. Then, for all $a \in [n_3]$, we have

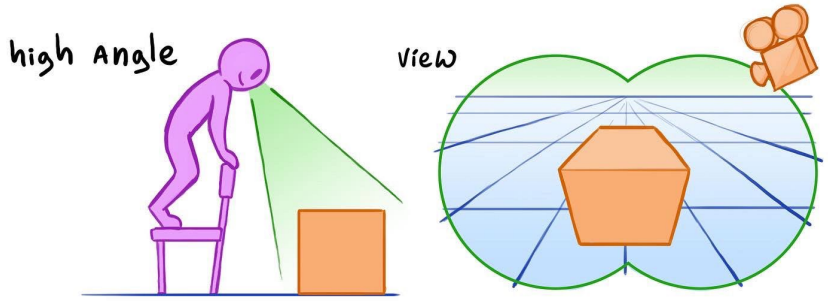
- $\text{colspan}\big(T(:,:,a)\big) \subseteq \text{span}\big(\{\boldsymbol{u}^{(i)}\}\big)$
- $\text{rowspan}\big(T(:,:,a)\big) \subseteq \text{span}\big(\{\boldsymbol{v}^{(i)}\}\big)$

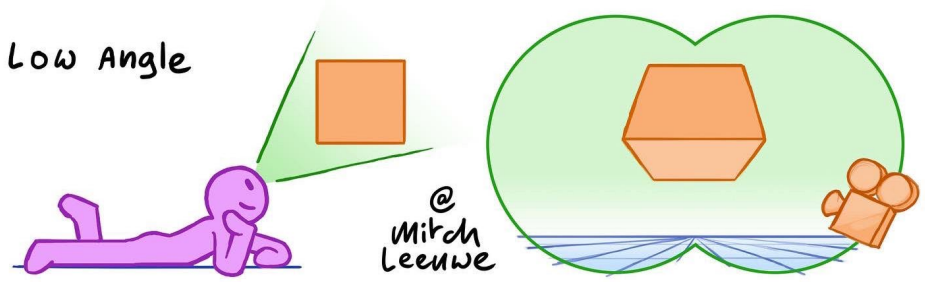
Detour: Tensor and Linear Algebra

- **Matrix:** single "view" of $\{\boldsymbol{u}^{(i)}\}$ and $\{\boldsymbol{v}^{(i)}\}$



- **Tensor:** multiple "views"





Detour: The Trouble with Tensors

Many features of matrices that we take for granted simply do not hold for tensors

- For a matrix M, $\mathrm{rank}(M) = \dim\big(\mathrm{colspan}(M)\big) = \dim\big(\mathrm{rowspan}(M)\big)$
- For a tensor, the dimension of the span of its columns may not equal to the dimension of the span of its rows
  - Consider a tensor T of size $(n_1, n_2, n_3) = (k^2, k, k)$ such that $\{T(i, :, :)\}_{i \in [k^2]}$ are the columns of the $k^2 \times k^2$ identity matrix
  - Then, $\dim(\mathrm{span}\{T(1, :, :), \dots, T(k^2, :, :)\}) = k^2$
  - However, $\dim(\mathrm{span}\{T(:, 1, :), \dots, T(:, k, :)\}) \leq k$

Detour: The Trouble with Tensors

Many features of matrices that we take for granted simply do not hold for tensors

- The rank of a matrix will not change if we move from $\mathbb{R}$ to $\mathbb{C}$
- Tensor rank depends on the field
  - Consider the following tensor

$$T = \left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\right] = \frac{1}{2}\begin{bmatrix} 1 \\ -i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -i \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 \\ i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \end{bmatrix}$$

  - $\text{rank}_{\mathbb{C}}(T) = 2$
  - $\text{rank}_{\mathbb{R}}(T) = 3$

Detour: The Trouble with Tensors

Many features of matrices that we take for granted simply do not hold for tensors

- If a matrix has rank r, then there is a limit of how well any rank $r' < r$ matrix can approximate it (Eckart–Young–Mirsky theorem)
- There are tensors of rank 3, but which are arbitrarily close to tensors of rank 2
  - Consider the following tensor

$$T = \left[\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right]$$

  - It can be approximated by

$$\left[\begin{bmatrix} 0 & 1 \\ 1 & 1/n \end{bmatrix}, \begin{bmatrix} 1 & 1/n \\ 1/n & 1/n^2 \end{bmatrix}\right] = \begin{bmatrix} 1 \\ 1/n \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1/n \end{bmatrix} \otimes \begin{bmatrix} n \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} n \\ 0 \end{bmatrix}$$

- The border rank of T is the smallest number r such that $\forall \epsilon > 0$, $\exists T'$ of rank $\leq r$ such that T' is entry-wise ϵ close to T

Detour: The Trouble with Tensors

Computationally, basic linear algebraic primitives are intractable for tensors.

- **Hillar-Lim '13:** Most tensor problems are **NP**-hard

Eigenvalue over $\mathbb{R}$	NP-hard (Theorem 1.3)
Approximating Eigenvector over $\mathbb{R}$	NP-hard (Theorem 1.5)
Symmetric Eigenvalue over $\mathbb{R}$	NP-hard (Theorem 9.3)
Approximating Symmetric Eigenvalue over $\mathbb{R}$	NP-hard (Theorem 9.6)
Singular Value over $\mathbb{R}$, $\mathbb{C}$	NP-hard (Theorem 1.7)
Symmetric Singular Value over $\mathbb{R}$	NP-hard (Theorem 10.2)
Approximating Singular Vector over $\mathbb{R}$, $\mathbb{C}$	NP-hard (Theorem 6.3)
Spectral Norm over $\mathbb{R}$	NP-hard (Theorem 1.10)
Symmetric Spectral Norm over $\mathbb{R}$	NP-hard (Theorem 10.2)
Approximating Spectral Norm over $\mathbb{R}$	NP-hard (Theorem 1.11)
Nonnegative Definiteness	NP-hard (Theorem 11.2)
Best Rank-1 Approximation	NP-hard (Theorem 1.13)
Best Symmetric Rank-1 Approximation	NP-hard (Theorem 10.2)
Rank over $\mathbb{R}$ or $\mathbb{C}$	NP-hard (Theorem 8.2)

Tensor Decomposition

Given a tensor $T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ such that

$$T = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)}$$

Our goal is to recover the set of factors $\left\{\left(\boldsymbol{u}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{w}^{(i)}\right)\right\}_{i \in [r]}$

- There are some symmetries in this decomposition
- $\left\{\left(\boldsymbol{u}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{w}^{(i)}\right)\right\}$ and $\left\{\left(\widehat{\boldsymbol{u}}^{(i)}, \widehat{\boldsymbol{v}}^{(i)}, \widehat{\boldsymbol{w}}^{(i)}\right)\right\}$ are equivalent if there exists a permutation $\pi : [r] \to [r]$ such that

$$\boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)} = \widehat{\boldsymbol{u}}^{(i)} \otimes \widehat{\boldsymbol{v}}^{(i)} \otimes \widehat{\boldsymbol{w}}^{(i)}, \qquad \forall\, i \in [r]$$

When are the factors of T determined up to equivalence?

Tensor Decomposition

Theorem (Harshman, Jennrich ’70**).** Suppose the following conditions hold:

1) $\{u_i\}$ are linearly independent
2) $\{v_i\}$ are linearly independent
3) $n_3 \geq 2$ and no two w_i, w_j are collinear

Then the factors are uniquely determined up to equivalence, and there is a polynomial time algorithm to find them

Tensor decomposition: Jennrich's algorithm

1. Choose $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{n_3}$ uniformly random unit vectors
2. Set

$$T(\boldsymbol{a}) \coloneqq \sum_{i \in [n_3]} a_i T(:,:,i) \quad \text{and} \quad T(\boldsymbol{b}) \coloneqq \sum_{i \in [n_3]} b_i T(:,:,i)$$

tensor contraction

3. Compute $A \coloneqq T(\boldsymbol{a})T(\boldsymbol{b})^{-1}$ and $B \coloneqq \left(T(\boldsymbol{a})^{-1}T(\boldsymbol{b})\right)^{\top}$
4. Let $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(r)}$ be eigenvectors of A with eigenvalues $\lambda_1, \ldots, \lambda_r$
5. Let $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(r)}$ be eigenvectors of B with eigenvalues $\lambda_1^{-1}, \ldots, \lambda_r^{-1}$
6. Solve linear system to recover $\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(r)}$:

$$T = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)}$$

Tensor decomposition: Jennrich's algorithm

- Let $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$ be the matrices with columns $\{\boldsymbol{u}^{(i)}\}$ and $\{\boldsymbol{v}^{(i)}\}$
- Let $D_{\boldsymbol{a}}, D_{\boldsymbol{b}} \in \mathbb{R}^{r \times r}$ be diagonal matrices with entries $\langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle$ and $\langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle$

Lemma. We have that

$$T(\boldsymbol{a}) = UD_{\boldsymbol{a}}V^\top, \qquad T(\boldsymbol{b}) = UD_{\boldsymbol{b}}V^\top$$

Proof.

$$\begin{aligned} T(\boldsymbol{a}) &= \sum_{i \in [n_3]} a_i T(:,:,i) = \sum_{i \in [n_3]} a_i \sum_{j \in [r]} \left(\boldsymbol{u}^{(j)} \otimes \boldsymbol{v}^{(j)}\right)\left(\boldsymbol{w}^{(j)}\right)_i \\ &= \sum_{j \in [r]} \boldsymbol{u}^{(j)} \otimes \boldsymbol{v}^{(j)} \langle a, \boldsymbol{w}^{(j)} \rangle \\ &= UD_{\boldsymbol{a}}V^\top \end{aligned}$$

■

Tensor decomposition: Jennrich's algorithm

- Let $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$ be the matrices with columns $\{\boldsymbol{u}^{(i)}\}$ and $\{\boldsymbol{v}^{(i)}\}$
- Let $D_{\boldsymbol{a}}, D_{\boldsymbol{b}} \in \mathbb{R}^{r \times r}$ be diagonal matrices with entries $\langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle$ and $\langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle$

Lemma. We have that

$$T(\boldsymbol{a}) = UD_{\boldsymbol{a}}V^\top, \qquad T(\boldsymbol{b}) = UD_{\boldsymbol{b}}V^\top$$

- Consider the following generalized eigenvalue problem:

$$T(\boldsymbol{a})\boldsymbol{v} = \lambda_{\boldsymbol{v}} T(\boldsymbol{b})\boldsymbol{v}, \qquad UD_{\boldsymbol{a}}V^\top \boldsymbol{v} = \lambda_{\boldsymbol{v}} UD_{\boldsymbol{b}}V^\top \boldsymbol{v}$$

- $\{\boldsymbol{v}^{(i)}\}$ are eigenvectors, suppose they are orthonormal. Then,

$$UD_{\boldsymbol{a}}V^\top \boldsymbol{v}^{(i)} = UD_{\boldsymbol{a}}\boldsymbol{e}_i = \langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle U\boldsymbol{e}_i = \langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle \boldsymbol{u}^{(i)}$$
$$UD_{\boldsymbol{b}}V^\top \boldsymbol{v}^{(i)} = UD_{\boldsymbol{b}}\boldsymbol{e}_i = \langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle U\boldsymbol{e}_i = \langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle \boldsymbol{u}^{(i)}$$

$$T(\boldsymbol{a})\boldsymbol{v}^{(i)} = \lambda_i T(\boldsymbol{b})\boldsymbol{v}^{(i)}, \qquad \lambda_i = \frac{\langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle}{\langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle}$$

Tensor decomposition: Jennrich's algorithm

- Let $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$ be the matrices with columns $\{\boldsymbol{u}^{(i)}\}$ and $\{\boldsymbol{v}^{(i)}\}$
- Let $D_{\boldsymbol{a}}, D_{\boldsymbol{b}} \in \mathbb{R}^{r \times r}$ be diagonal matrices with entries $\langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle$ and $\langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle$

$$T(\boldsymbol{a})\boldsymbol{v}^{(i)} = \lambda_i T(\boldsymbol{b})\boldsymbol{v}^{(i)}, \qquad \lambda_i = \frac{\langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle}{\langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle}$$

- If all λ_i are distinct, then solving this generalized eigenvalue problem can uniquely recover the generalized eigenvectors $\{\boldsymbol{v}^{(i)}\}$ (up to a rescaling)

Tensor decomposition: Jennrich's algorithm

Lemma. Almost surely, the columns of U and V are the unique eigenvectors corresponding to non-zero eigenvalues of $T(\boldsymbol{a})T(\boldsymbol{b})^{-1}$ and $\left(T(\boldsymbol{a})^{-1}T(\boldsymbol{b})\right)^{\top}$ respectively. Moreover, the eigenvalue corresponding to $\boldsymbol{u}^{(i)}$ is the reciprocal of the eigenvalue corresponding to $\boldsymbol{v}^{(i)}$

Proof.

- Using the formula for $T(\boldsymbol{a})$ and $T(\boldsymbol{b})$, we have

$$T(\boldsymbol{a})T(\boldsymbol{b})^{-1} = UD_{\boldsymbol{a}}V^{\top}(UD_{\boldsymbol{b}}V^{\top})^{-1} = UD_{\boldsymbol{a}}V^{\top}(V^{\top})^{-1}D_{\boldsymbol{b}}^{-1}U^{-1} = UD_{\boldsymbol{a}}D_{\boldsymbol{b}}^{-1}U^{-1}$$

$$\left(T(\boldsymbol{a})^{-1}T(\boldsymbol{b})\right)^{\top} = \left((V^{\top})^{-1}D_{\boldsymbol{a}}^{-1}U^{-1}UD_{\boldsymbol{b}}V^{\top}\right)^{\top} = \left((V^{\top})^{-1}D_{\boldsymbol{a}}^{-1}D_{\boldsymbol{b}}V^{\top}\right)^{\top} = VD_{\boldsymbol{b}}D_{\boldsymbol{a}}^{-1}V^{-1}$$

- The diagonal entries of $D_{\boldsymbol{a}}D_{\boldsymbol{b}}^{-1}$ is $\langle \boldsymbol{a}, \boldsymbol{w}^{(i)} \rangle / \langle \boldsymbol{b}, \boldsymbol{w}^{(i)} \rangle$
- Since every pair of vectors in $\{\boldsymbol{w}^{(i)}\}$ are not collinear and $\boldsymbol{a}, \boldsymbol{b}$ are random vectors, almost surely all diagonal entries are non-zero and distinct
- Thus, $\{\boldsymbol{u}^{(i)}\}$ and $\{\boldsymbol{v}^{(i)}\}$ are unique eigenvectors, and they are matched by reciprocal eigenvalues

■

Tensor decomposition: Jennrich's algorithm

- To recover $w^{(i)}$, we solve a linear system:

$$T = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)}$$

- The ground-truth $\{w^{(i)}\}$ is a solution
- We need to show that this is the only solution
- Observe that this big linear system can be decomposed as n_3 independent smaller linear systems:

$$T(:,:,a) = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} w_a^{(i)}, \qquad a \in [n_3]$$

$$[\boldsymbol{u}^{(1)} \otimes \boldsymbol{v}^{(1)} \quad \cdots \quad \boldsymbol{u}^{(r)} \otimes \boldsymbol{v}^{(r)}] \begin{bmatrix} w_a^{(1)} \\ \vdots \\ w_a^{(r)} \end{bmatrix} = T(:,:,a)$$

Tensor decomposition: Jennrich's algorithm

Lemma. $\{\boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)}\}_{i\in[r]}$ are linearly independent

Proof.

- Suppose there exists not all zero $\alpha_1, \ldots, \alpha_r$ such that

$$\sum_{i=1}^{r} \alpha_i \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} = \boldsymbol{0}$$

- Wlog, assume $\alpha_1 \neq 0$. Since $\{\boldsymbol{v}^{(i)}\}$ are linearly independent, we can find $a \in \mathbb{R}^{n_2}$ such that

$$\langle \boldsymbol{a}, \boldsymbol{v}^{(1)} \rangle \neq 0, \qquad \langle \boldsymbol{a}, \boldsymbol{v}^{(i)} \rangle = 0 \quad \forall i \neq 1$$

$$\boldsymbol{0} = \sum_{i=1}^{r} \alpha_i \boldsymbol{u}^{(i)} \otimes \langle \boldsymbol{a}, \boldsymbol{v}^{(i)} \rangle = \alpha_1 \boldsymbol{u}^{(1)} \langle \boldsymbol{a}, \boldsymbol{v}^{(1)} \rangle$$

- This is a contradiction since RHS $\neq 0$

■

The Method of Moments

Third moment ($d \times d \times d$ tensor)

Topic model

$$\Pr[x_1, x_2] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$\Pr[x_1, x_2, x_3] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

- Form the tensor $\tilde{T} \in \mathbb{R}^{d \times d \times d}$
- Run Jennrich's algorithm to recover w_i and $\boldsymbol{\mu}_i$

The Method of Moments

Third moment ($d \times d \times d$ tensor)

GMM

$$\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}] = \mathbb{E}[(\boldsymbol{\mu}_i + \boldsymbol{\eta}) \otimes (\boldsymbol{\mu}_i + \boldsymbol{\eta}) \otimes (\boldsymbol{\mu}_i + \boldsymbol{\eta})]$$

$$= \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i^{\otimes 3}$$

$$+ \sum_{i=1}^{k} w_i \sigma^2 \sum_{a=1}^{d} \boldsymbol{\mu}_i \otimes \boldsymbol{e}_a \otimes \boldsymbol{e}_a + \boldsymbol{e}_a \otimes \boldsymbol{e}_a \otimes \boldsymbol{\mu}_i + \boldsymbol{e}_a \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{e}_a$$

$$\sum_{i=1}^{k} w_i \boldsymbol{\mu}_i^{\otimes 3} = \mathbb{E}\left[\boldsymbol{x}^{\otimes 3} - \sigma^2 \sum_{a=1}^{d} \boldsymbol{x} \otimes \boldsymbol{e}_a \otimes \boldsymbol{e}_a + \boldsymbol{e}_a \otimes \boldsymbol{e}_a \otimes \boldsymbol{x} + \boldsymbol{e}_a \otimes \boldsymbol{x} \otimes \boldsymbol{e}_a\right]$$

The Method of Moments

Third moment ($d \times d \times d$ tensor)

GMM

$$\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i + \sigma^2 I$$

- Let $\boldsymbol{m} = \mathbb{E}[\boldsymbol{x}]$
- The covariance matrix of $\boldsymbol{x}$ is

$$\mathrm{Cov}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^\top] = \underbrace{\sum_{i=1}^{k} w_i (\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^\top}_{\succcurlyeq 0,\ \mathrm{rank} \le k-1} + \sigma^2 I$$

- Thus, $\lambda_{\min}\big(\mathrm{Cov}(\boldsymbol{x})\big) = \sigma^2$

The Method of Moments

If the σ_i^2 are different,

$$\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i + \overline{\sigma}^2 I, \qquad \overline{\sigma}^2 = \sum_{i=1}^{k} w_i \sigma_i^2$$

$$\mathrm{Cov}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^\top] = \sum_{i=1}^{k} w_i (\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^\top + \overline{\sigma}^2 I$$

- Let $\boldsymbol{v}$ be the smallest eigenvector of $\mathrm{Cov}(x)$ with $\|\boldsymbol{v}\| = 1$
- Then, $\langle \boldsymbol{v}, \boldsymbol{\mu}_i - \boldsymbol{m} \rangle = 0$ for every $i \in [k]$. We have

$$\mathbb{E}[\boldsymbol{x}\langle \boldsymbol{v}, \boldsymbol{x} - \boldsymbol{m}\rangle^2 | i] = \mathbb{E}[(\boldsymbol{\mu_i} + \sigma_i \boldsymbol{\eta})\langle \boldsymbol{v}, \boldsymbol{\mu_i} + \sigma_i \boldsymbol{\eta} - \boldsymbol{m}\rangle^2 | i] = \mathbb{E}[(\boldsymbol{\mu_i} + \sigma_i \boldsymbol{\eta})\langle \boldsymbol{v}, \sigma_i \boldsymbol{\eta}\rangle^2 | i]$$
$$= \sigma_i^2 \boldsymbol{\mu_i} \mathbb{E}[\boldsymbol{v}^\top \boldsymbol{\eta}\boldsymbol{\eta}^\top \boldsymbol{v} | i] = \sigma_i^2 \boldsymbol{\mu_i}$$

$$\mathbb{E}[\boldsymbol{x}\langle \boldsymbol{v}, \boldsymbol{x} - \boldsymbol{m}\rangle^2] = \sum_{i=1}^{k} w_i \sigma_i^2 \boldsymbol{\mu}_i$$

The Method of Moments

$$M_1 \coloneqq \mathbb{E}[\boldsymbol{x}\langle \boldsymbol{v}, \boldsymbol{x}-\boldsymbol{m}\rangle^2] = \sum_{i=1}^{k} w_i \mathbb{E}[\boldsymbol{x}\langle \boldsymbol{v}, \boldsymbol{x}-\boldsymbol{m}\rangle^2 | i] = \sum_{i=1}^{k} w_i \sigma_i^2 \boldsymbol{\mu}_i$$

$$M_2 \coloneqq \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] - \overline{\sigma}^2 I = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$M_3 \coloneqq \mathbb{E}\left[\boldsymbol{x}^{\otimes 3}\right] - \sum_{a=1}^{d} M_1 \otimes \boldsymbol{e}_a \otimes \boldsymbol{e}_a + \boldsymbol{e}_a \otimes \boldsymbol{e}_a \otimes M_1 + \boldsymbol{e}_a \otimes M_1 \otimes \boldsymbol{e}_a = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i^{\otimes 3}$$

Jennrich's Algorithm

Theorem. Let $T = \sum_{i=1}^{r} \boldsymbol{u}^{(i)} \otimes \boldsymbol{v}^{(i)} \otimes \boldsymbol{w}^{(i)} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with $\left\|\boldsymbol{w}^{(i)}\right\| = 1$. Suppose

- $U = \left[\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}, \ldots, \boldsymbol{u}^{(r)}\right] \in \mathbb{R}^{n_1 \times r}$ is full rank and has condition number $\kappa(U) := \frac{\sigma_{\max}(U)}{\sigma_{\min}(U)} \leq \kappa$
- $V = \left[\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(r)}\right] \in \mathbb{R}^{n_2 \times r}$ is full rank and has condition number $\kappa(V) \leq \kappa$
- $\left\|\boldsymbol{w}^{(i)} \pm \boldsymbol{w}^{(j)}\right\| \geq \delta$ for any $i \neq j \in [r]$

Let $\tilde{T} = T + E \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with

$$\|E\|_{\max} \leq \frac{\epsilon \|T\|_F}{\mathrm{poly}(\kappa, n, \delta^{-1}, \eta^{-1})}, \qquad n := \max\{n_1, n_2, n_3\}$$

Then, Jennrich's algorithm can output $\left\{\left(\widetilde{\boldsymbol{u}}^{(i)}, \widetilde{\boldsymbol{v}}^{(i)}, \widetilde{\boldsymbol{w}}^{(i)}\right)\right\}_{\boldsymbol{i} \in [\boldsymbol{r}]}$ such that with probability $1 - \eta$,

$$\left\|\widetilde{\boldsymbol{u}}^{(i)} \otimes \widetilde{\boldsymbol{v}}^{(i)} \otimes \widetilde{\boldsymbol{w}}^{(i)} - \boldsymbol{u}^{(\pi(i))} \otimes \boldsymbol{v}^{(\pi(i))} \otimes \boldsymbol{w}^{(\pi(i))}\right\|_F \leq \epsilon \|T\|_F, \qquad \forall i \in [k]$$

for some permutation π