Tensor Decompositions and Their Applications

Ankur Moitra (MIT)

IPAM Tutorial, Part 1

SPEARMAN'S HYPOTHESIS

Charles Spearman (1904): There are two types of intelligence, *eductive* and *reproductive*

SPEARMAN'S HYPOTHESIS

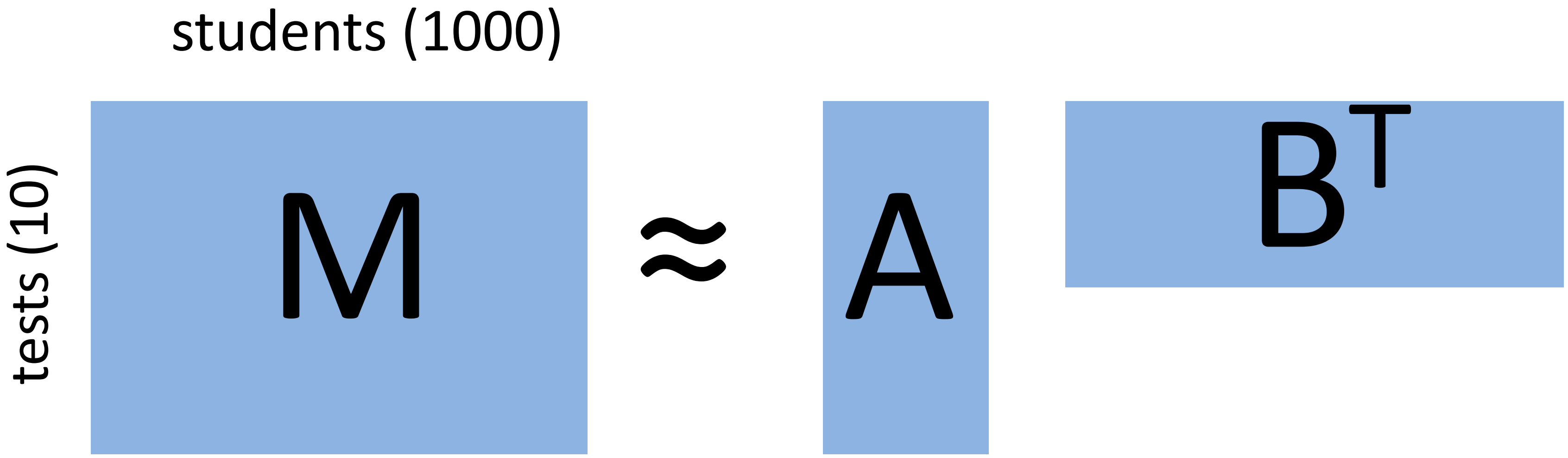
Charles Spearman (1904): There are two types of intelligence, *eductive* and *reproductive*

eductive (adj): the ability to make sense out of complexity
reproductive (adj): the ability to store and reproduce information

SPEARMAN'S HYPOTHESIS

Charles Spearman (1904): There are two types of intelligence, *eductive* and *reproductive*

To test this theory, he invented **Factor Analysis:**

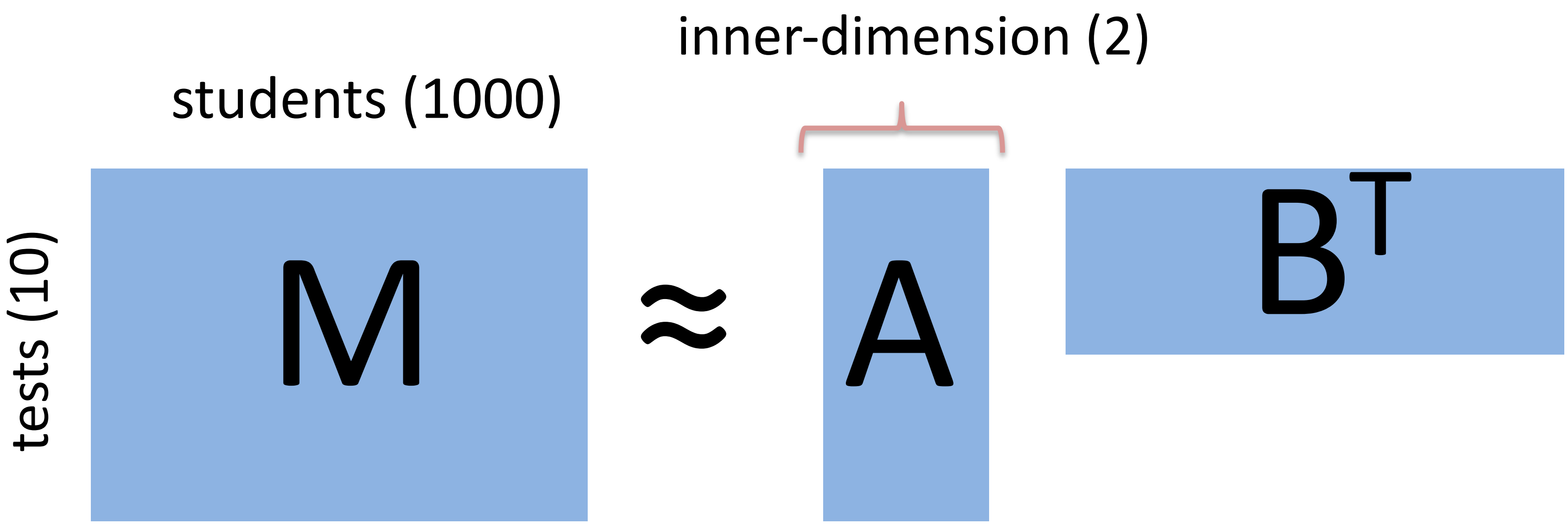


eductive (adj): the ability to make sense out of complexity
reproductive (adj): the ability to store and reproduce information

SPEARMAN'S HYPOTHESIS

Charles Spearman (1904): There are two types of intelligence, *eductive* and *reproductive*

To test this theory, he invented **Factor Analysis:**



eductive (adj): the ability to make sense out of complexity
reproductive (adj): the ability to store and reproduce information

Given: $$M = \sum a_i \otimes b_i$$

$$= \underbrace{AB^\top}_{\text{“correct” factors}}$$

Given: $$M = \sum a_i \otimes b_i$$
$$= \underbrace{AB^\top}_{\text{"correct" factors}}$$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Given: $M = \sum a_i \otimes b_i$

$$= \underbrace{AB^\top}_{\text{“correct” factors}} = \underbrace{\left(AR\right)\left(R^{-1}B^\top\right)}_{\text{alternative factorization}}$$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Given: $$M = \sum a_i \otimes b_i$$

$$= \underbrace{AB^\top}_{\text{“correct” factors}} = \underbrace{\left(AR\right)\left(R^{-1}B^\top\right)}_{\text{alternative factorization}}$$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Claim: The factors $\{a_i\}$ and $\{b_i\}$ are not determined uniquely unless we impose additional conditions on them

Given: $M = \sum a_i \otimes b_i$

$$= \underbrace{AB^\top}_{\text{"correct" factors}} = \underbrace{\big(AR\big)\big(R^{-1}B^\top\big)}_{\text{alternative factorization}}$$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Claim: The factors $\{a_i\}$ and $\{b_i\}$ are not determined uniquely unless we impose additional conditions on them

e.g. if $\{a_i\}$ and $\{b_i\}$ are orthogonal, or $\mathrm{rank}(M) = 1$

Given: $M = \sum a_i \otimes b_i$

$$= \underbrace{AB^\top}_{\text{“correct” factors}} = \underbrace{\left(AR\right)\left(R^{-1}B^\top\right)}_{\text{alternative factorization}}$$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Claim: The factors $\{a_i\}$ and $\{b_i\}$ are not determined uniquely unless we impose additional conditions on them

e.g. if $\{a_i\}$ and $\{b_i\}$ are orthogonal, or $\mathrm{rank}(M) = 1$

This is called the **rotation problem**, and is a major issue in factor analysis and motivates the study of **tensor methods**…

OUTLINE

Part I: Introduction

- The Rotation Problem
- Jennrich's Algorithm

Part II: Applications

- Phylogenetic Reconstruction
- Mixtures of Gaussians
- Orbit Retrieval

OUTLINE

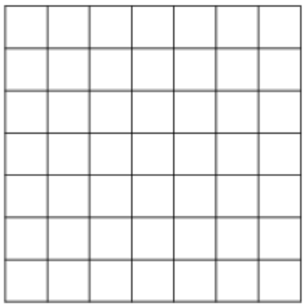
Part I: Introduction

- The Rotation Problem
- **Jennrich's Algorithm**

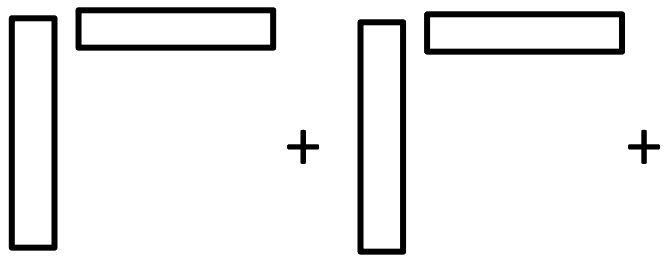
Part II: Applications

- Phylogenetic Reconstruction
- Mixtures of Gaussians
- Orbit Retrieval

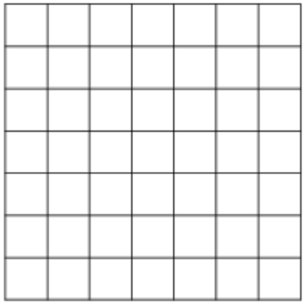
MATRIX DECOMPOSITIONS



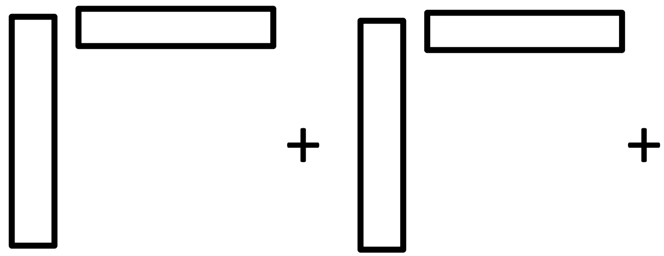
$$M = a_1 \otimes b_1 + a_2 \otimes b_2 + \cdots + a_R \otimes b_R$$



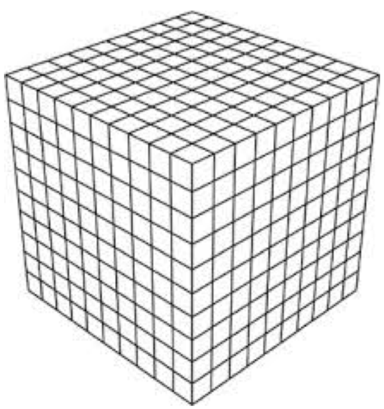
MATRIX DECOMPOSITIONS



$$M = a_1 \otimes b_1 + a_2 \otimes b_2 + \cdots + a_R \otimes b_R$$



TENSOR DECOMPOSITIONS



$$T = a_1 \otimes b_1 \otimes c_1 + \cdots + a_R \otimes b_R \otimes c_R$$

(i, j, k) entry of $x \otimes y \otimes z$ is $x(i) \times y(j) \times z(k)$

When are tensor decompositions unique?

When are tensor decompositions unique?

Theorem [Jennrich 1970]: Suppose $\{a_i\}$ and $\{b_i\}$ are linearly independent and no pair of vectors in $\{c_i\}$ is a scalar multiple of each other…

When are tensor decompositions unique?

Theorem [Jennrich 1970]: Suppose $\{a_i\}$ and $\{b_i\}$ are linearly independent and no pair of vectors in $\{c_i\}$ is a scalar multiple of each other. Then

$$T = a_1 \otimes b_1 \otimes c_1 + \cdots + a_R \otimes b_R \otimes c_R$$

is unique up to permuting the rank one terms and rescaling the factors.

When are tensor decompositions unique?

Theorem [Jennrich 1970]: Suppose $\{a_i\}$ and $\{b_i\}$ are linearly independent and no pair of vectors in $\{c_i\}$ is a scalar multiple of each other. Then

$$T = a_1 \otimes b_1 \otimes c_1 + \cdots + a_R \otimes b_R \otimes c_R$$

is unique up to permuting the rank one terms and rescaling the factors.

Equivalently, the rank one factors are **unique**

When are tensor decompositions unique?

Theorem [Jennrich 1970]: Suppose $\{a_i\}$ and $\{b_i\}$ are linearly independent and no pair of vectors in $\{c_i\}$ is a scalar multiple of each other. Then

$$T = a_1 \otimes b_1 \otimes c_1 + \cdots + a_R \otimes b_R \otimes c_R$$

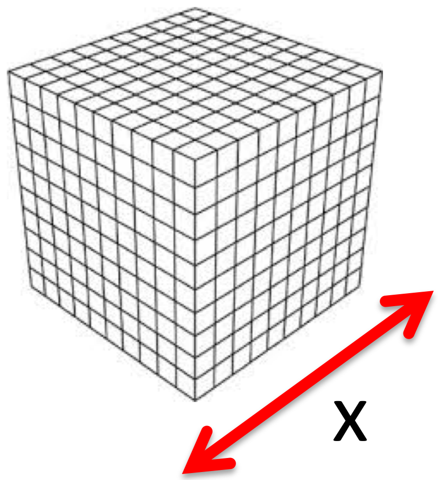
is unique up to permuting the rank one terms and rescaling the factors.

Equivalently, the rank one factors are **unique**

There is a simple algorithm to compute the factors too!

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x)$

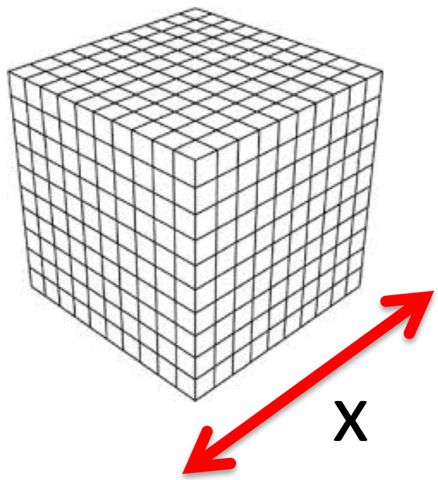


i.e. add up matrix slices

$$\sum_i x_i T_i$$

JENNRICH'S ALGORITHM

- Compute $T(\cdot,\cdot,x)$



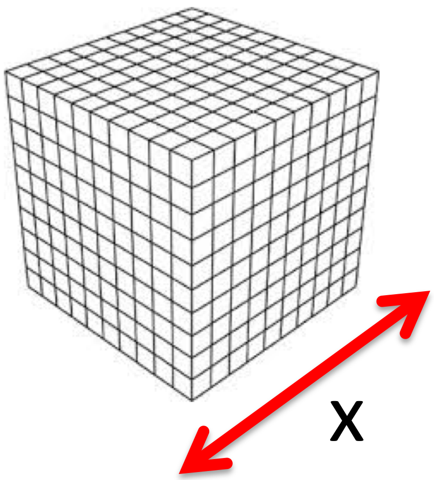
i.e. add up matrix slices

$$\sum_i x_i T_i$$

If $T = a \otimes b \otimes c$ then $T(\cdot,\cdot,x) = \langle c, x\rangle a \otimes b$

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x) = \sum \langle c_i, x \rangle a_i \otimes b_i$

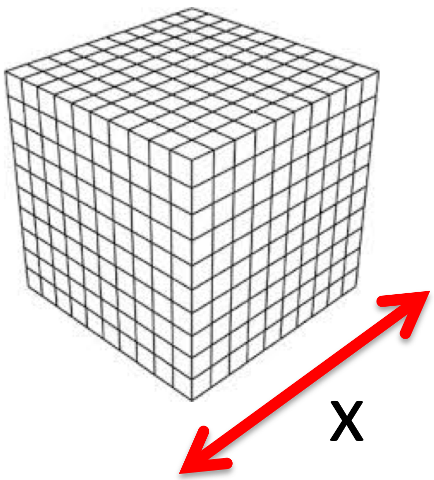


i.e. add up matrix slices

$$\sum_i x_i T_i$$

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x) = \sum \langle c_i, x \rangle a_i \otimes b_i$



i.e. add up matrix slices

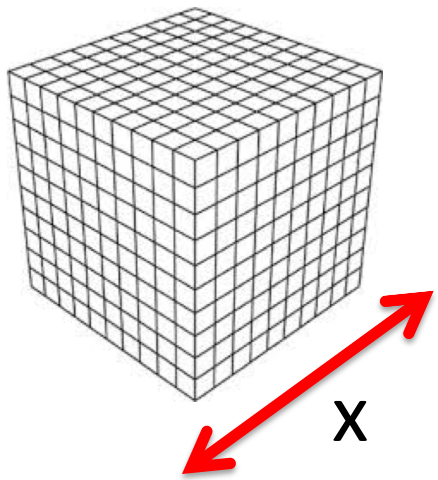
$$\sum_i x_i T_i$$

(x is chosen uniformly at random from $\mathbb{S}^{n-1}$)

JENNRICH'S ALGORITHM

$\mathrm{Diag}(\{\langle c_i, x\rangle\}_i)$

Compute $T(\cdot,\cdot,x) = AD_xB^\top$



i.e. add up matrix slices

$$\sum_i x_i T_i$$

(x is chosen uniformly at random from $\mathbb{S}^{n-1}$)

JENNRICH'S ALGORITHM

- Compute $T(\cdot,\cdot,x) = AD_xB^\top$

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x) = AD_x B^\top$
- Compute $T(\cdot, \cdot, y) = AD_y B^\top$

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x) = AD_xB^\top$
- Compute $T(\cdot, \cdot, y) = AD_yB^\top$
- Diagonalize $T(\cdot, \cdot, x)\Big(T(\cdot, \cdot, y)\Big)^{-1}$

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x) = AD_xB^\top$
- Compute $T(\cdot, \cdot, y) = AD_yB^\top$
- Diagonalize $\underbrace{T(\cdot, \cdot, x)\Big(T(\cdot, \cdot, y)\Big)^{-1}}_{AD_xB^\top (B^\top)^{-1} D_y^{-1} A^{-1}}$

JENNRICH'S ALGORITHM

- Compute $T(\cdot,\cdot,x) = AD_xB^\top$
- Compute $T(\cdot,\cdot,y) = AD_yB^\top$
- Diagonalize $\underbrace{T(\cdot,\cdot,x)\Big(T(\cdot,\cdot,y)\Big)^{-1}}_{AD_xD_y^{-1}A^{-1}}$

JENNRICH'S ALGORITHM

- Compute $T(\cdot,\cdot,x) = AD_xB^\top$
- Compute $T(\cdot,\cdot,y) = AD_yB^\top$
- Diagonalize $\underbrace{T(\cdot,\cdot,x)\Big(T(\cdot,\cdot,y)\Big)^{-1}}_{AD_xD_y^{-1}A^{-1}}$

Claim: whp (over x,y) the eigenvalues are distinct, so the Eigendecomposition is unique and recovers a_i

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x) = AD_x B^\top$
- Compute $T(\cdot, \cdot, y) = AD_y B^\top$
- Diagonalize $T(\cdot, \cdot, x)\Big(T(\cdot, \cdot, y)\Big)^{-1}$

JENNRICH'S ALGORITHM

- Compute $T(\cdot,\cdot,x) = AD_xB^\top$
- Compute $T(\cdot,\cdot,y) = AD_yB^\top$
- Diagonalize $T(\cdot,\cdot,x)\Big(T(\cdot,\cdot,y)\Big)^{-1}$
- Diagonalize $T(\cdot,\cdot,y)\Big(T(\cdot,\cdot,x)\Big)^{-1}$

JENNRICH'S ALGORITHM

- Compute $T(\cdot, \cdot, x) = AD_xB^\top$
- Compute $T(\cdot, \cdot, y) = AD_yB^\top$
- Diagonalize $T(\cdot, \cdot, x)\Big(T(\cdot, \cdot, y)\Big)^{-1}$
- Diagonalize $T(\cdot, \cdot, y)\Big(T(\cdot, \cdot, x)\Big)^{-1}$
- Match up the factors (their eigenvalues are reciprocals) and find $\{c_i\}_i$ by solving a linear syst.

Given: $M = \sum a_i \otimes b_i$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Only possible if $\{a_i\}$ and $\{b_i\}$ are orthogonal, or $\mathrm{rank}(M) = 1$

Given: $M = \sum a_i \otimes b_i$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Only possible if $\{a_i\}$ and $\{b_i\}$ are orthogonal, or $\mathrm{rank}(M) = 1$

Given: $T = \sum a_i \otimes b_i \otimes c_i$

When can we find the factors $\{a_i\}$, $\{b_i\}$ and $\{c_i\}$ uniquely?

Given: $M = \sum a_i \otimes b_i$

When can we find the factors $\{a_i\}$ and $\{b_i\}$ uniquely?

Only possible if $\{a_i\}$ and $\{b_i\}$ are orthogonal, or $\mathrm{rank}(M) = 1$

Given: $T = \sum a_i \otimes b_i \otimes c_i$

When can we find the factors $\{a_i\}$, $\{b_i\}$ and $\{c_i\}$ uniquely?

Jennrich: If $\{a_i\}$ and $\{b_i\}$ are full rank and no pair in $\{c_i\}$ are scalar multiples of each other

OUTLINE

Part I: Introduction

- The Rotation Problem
- Jennrich's Algorithm

Part II: Applications

- Phylogenetic Reconstruction
- Mixtures of Gaussians
- Orbit Retrieval

OUTLINE

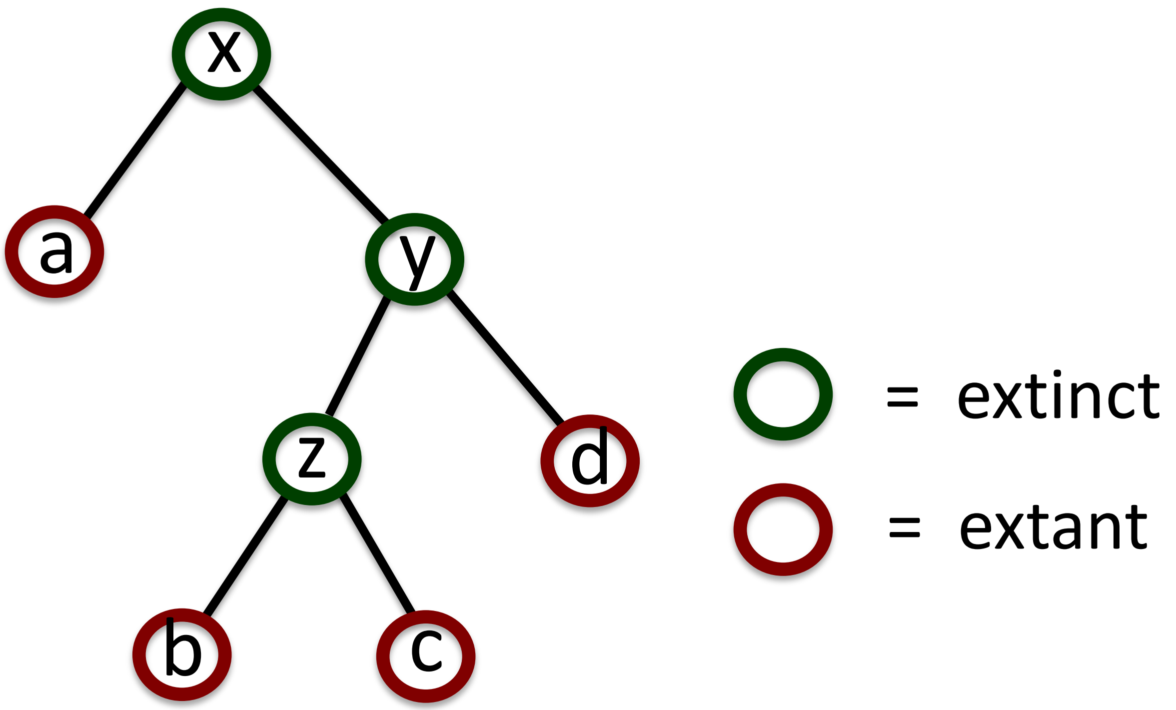
Part I: Introduction

- The Rotation Problem
- Jennrich's Algorithm

Part II: Applications

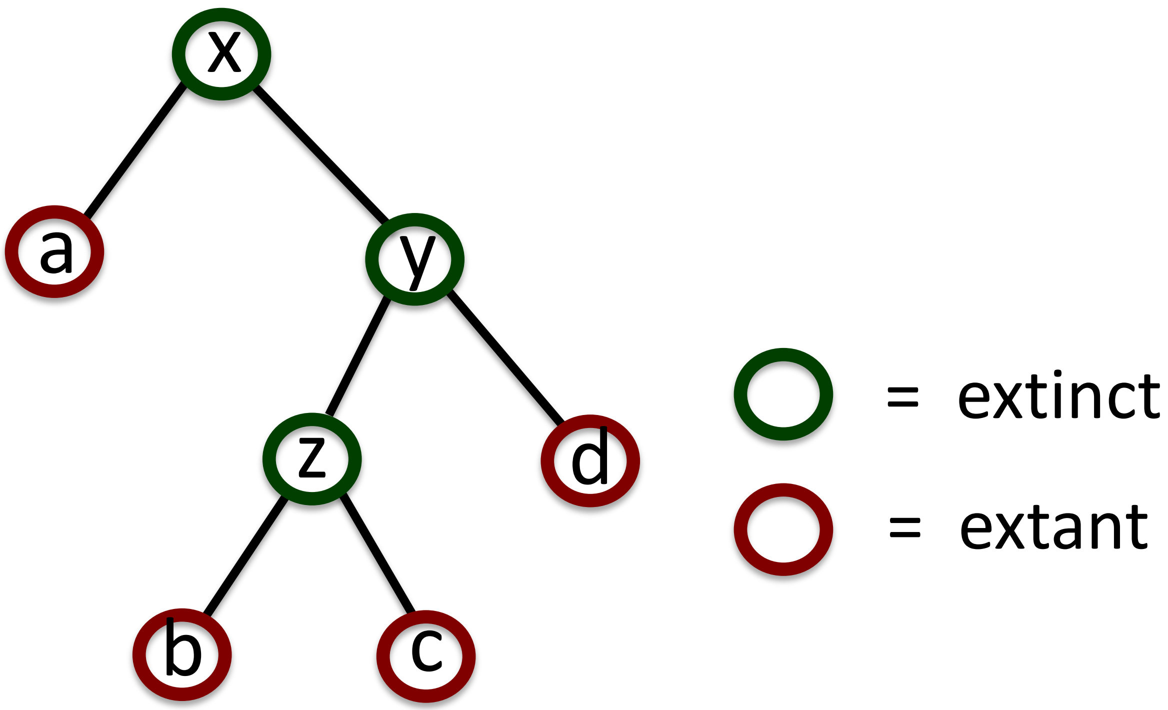
- **Phylogenetic Reconstruction**
- Mixtures of Gaussians
- Orbit Retrieval

PHYLOGENETIC RECONSTRUCTION

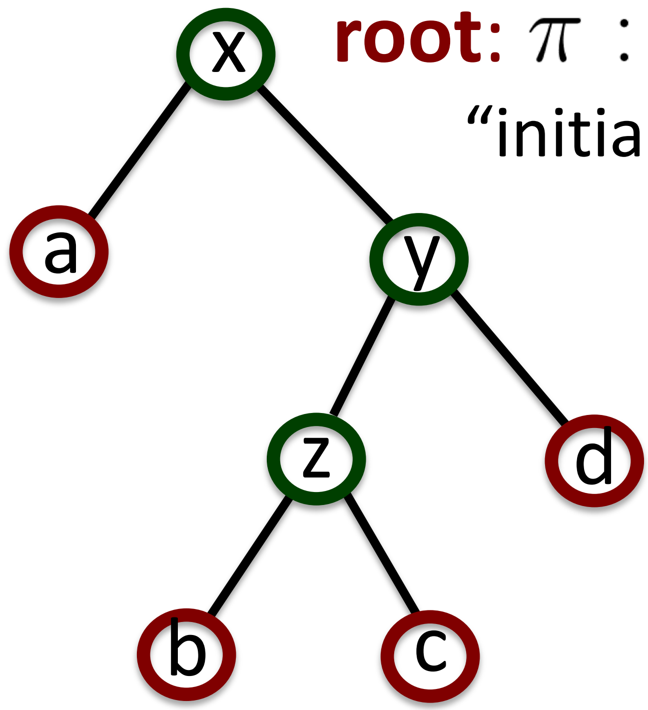


"Tree of Life"

PHYLOGENETIC RECONSTRUCTION

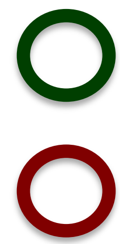


PHYLOGENETIC RECONSTRUCTION



root: $\pi : \Sigma \rightarrow \mathbb{R}^+$

"initial distribution"



= extinct

= extant

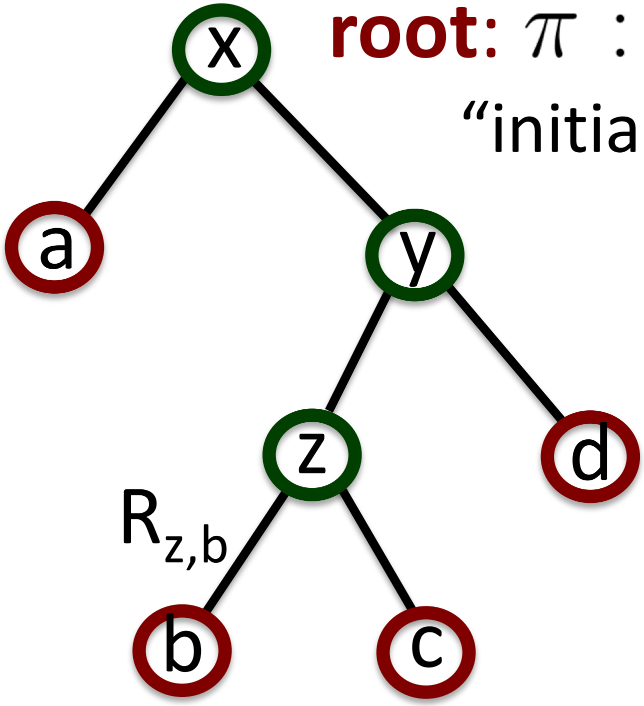
Σ = alphabet

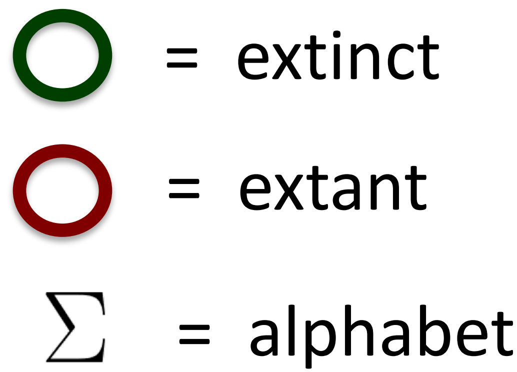
PHYLOGENETIC RECONSTRUCTION

root: $\pi : \Sigma \rightarrow \mathbb{R}^{+}$

"initial distribution"

"conditional distribution"



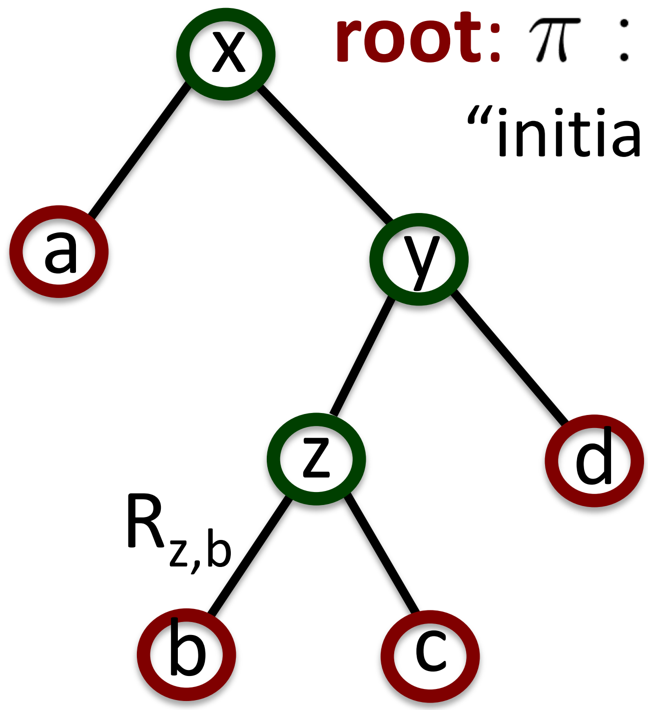


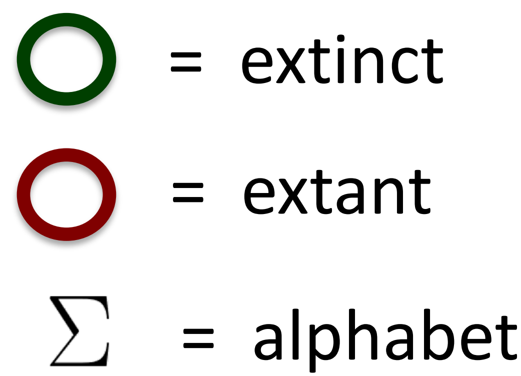
PHYLOGENETIC RECONSTRUCTION

root: $\pi : \Sigma \rightarrow \mathbb{R}^{+}$

"initial distribution"

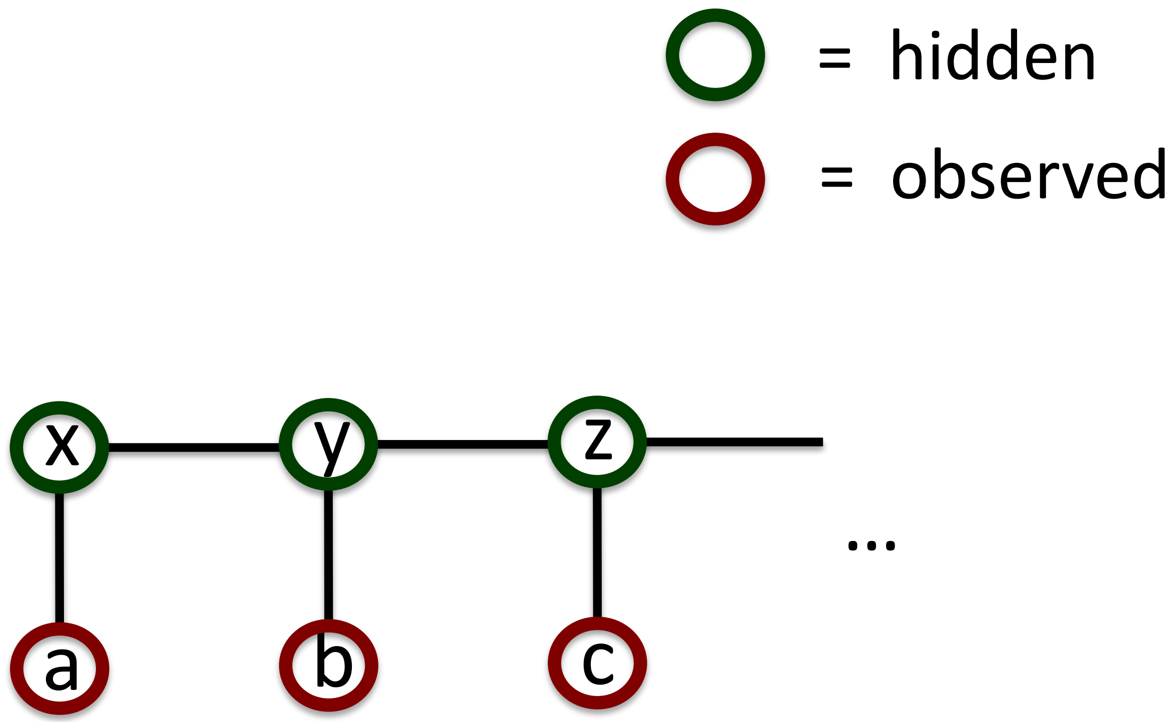
"conditional distribution"





In each sample, we observe a symbol (Σ) at each extant ( ) node where we sample from π for the root, and propagate it using $R_{x,y}$, etc

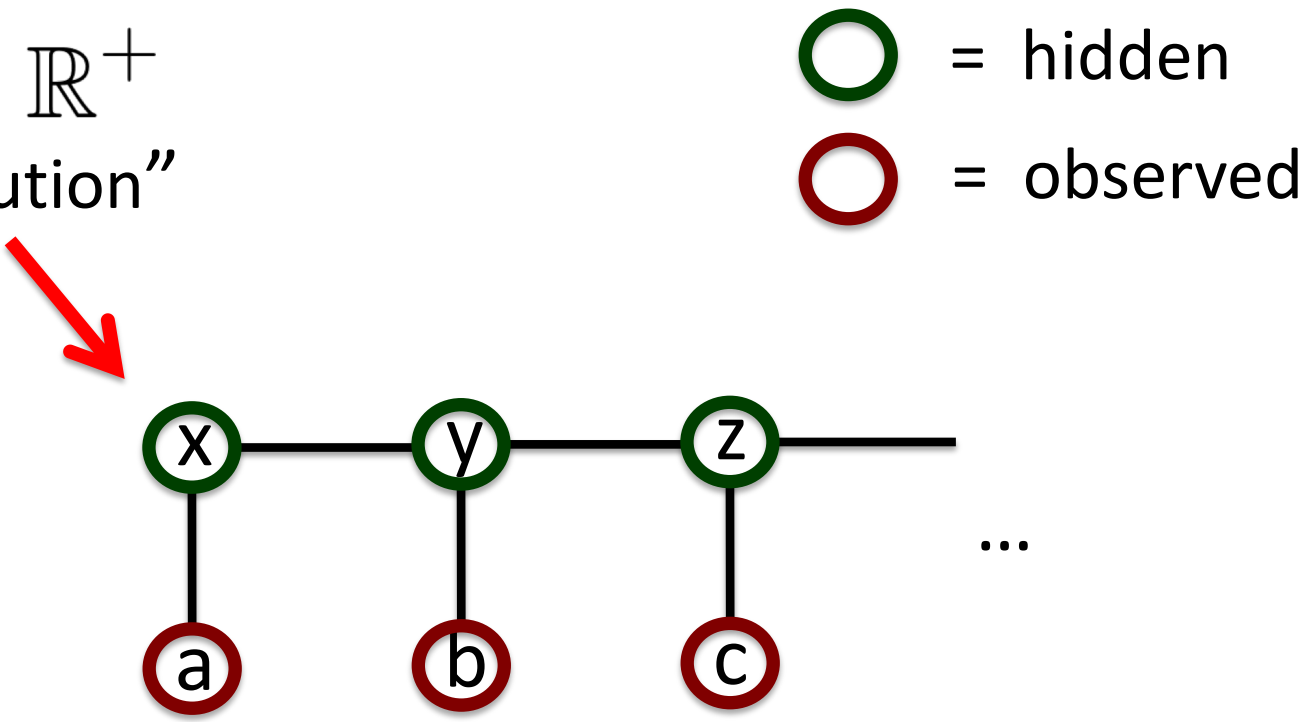
HIDDEN MARKOV MODELS



HIDDEN MARKOV MODELS

$$\pi : \Sigma_S \to \mathbb{R}^+$$

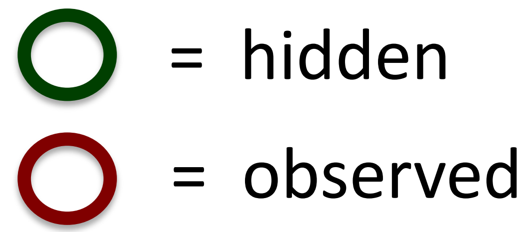
"initial distribution"

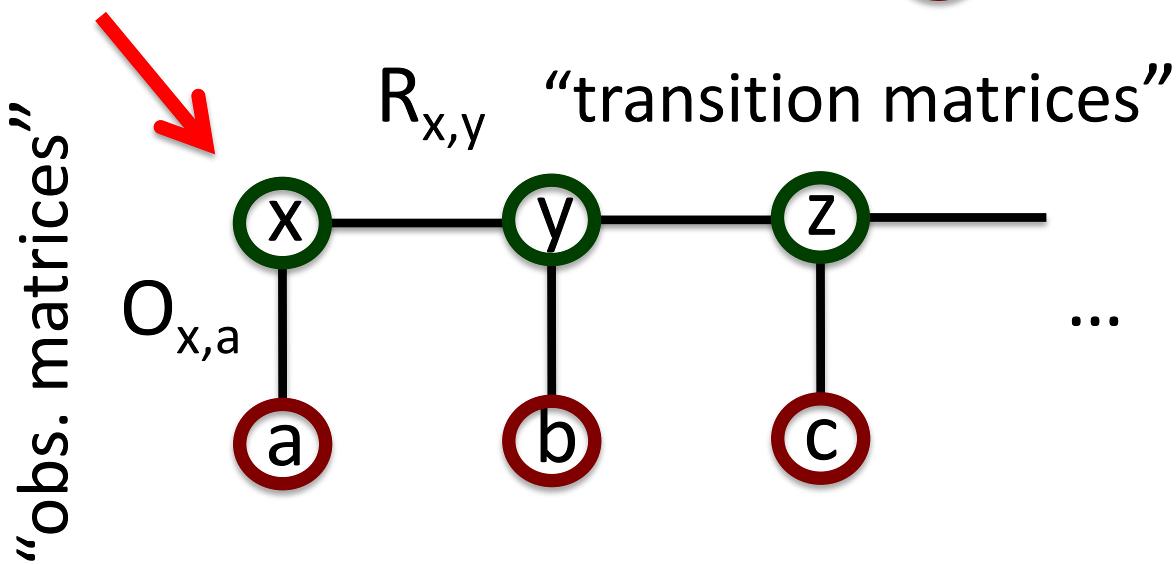


HIDDEN MARKOV MODELS

$$\pi : \Sigma_S \to \mathbb{R}^+$$

"initial distribution"

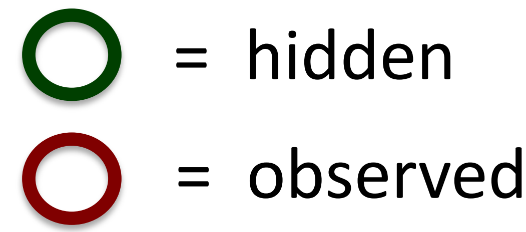


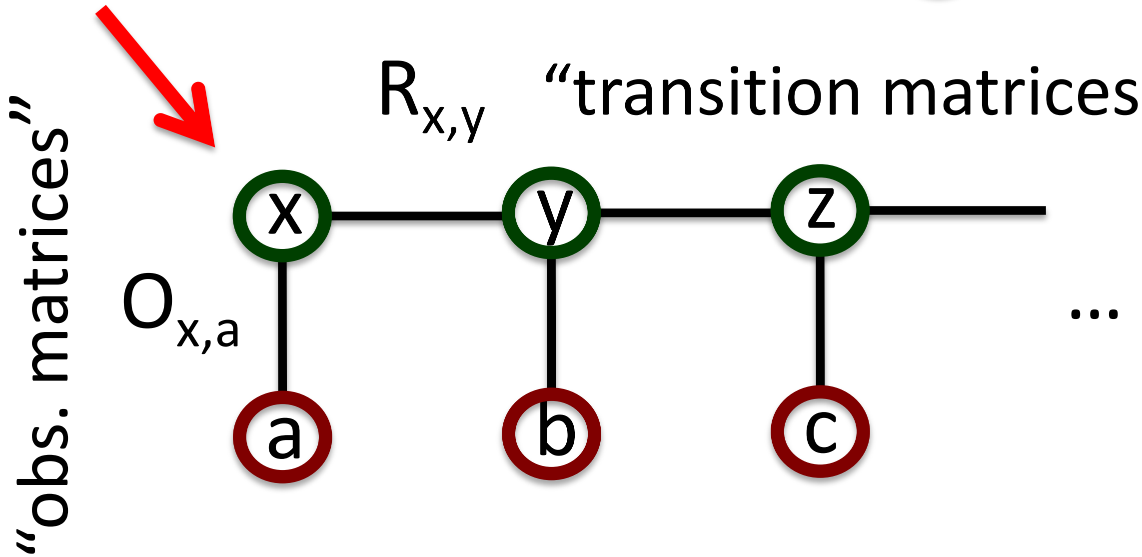


HIDDEN MARKOV MODELS

$$\pi : \Sigma_S \to \mathbb{R}^+$$

"initial distribution"





In each sample, we observe a symbol (Σ_O) at each obs. ( ) node where we sample from π for the start, and propagate it using $R_{x,y}$, etc (Σ_S)

Can we reconstruct just the topology from random samples?

Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

[Steel, 1994]: The following is a distance function on the edges

$$d_{x,y} = -\ln|\det(P_{x,y})| + \frac{1}{2}\prod_{\sigma \text{ in } \Sigma} \pi_{x,\sigma} - \frac{1}{2}\prod_{\sigma \text{ in } \Sigma} \pi_{y,\sigma}$$

where $P_{x,y}$ is the joint distribution

Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

[Steel, 1994]: The following is a distance function on the edges

$$d_{x,y} = -\ln|\det(P_{x,y})| + \frac{1}{2}\prod_{\sigma \text{ in } \Sigma} \pi_{x,\sigma} - \frac{1}{2}\prod_{\sigma \text{ in } \Sigma} \pi_{y,\sigma}$$

where $P_{x,y}$ is the joint distribution, and the distance between leaves is the sum of distances on the path in the tree

Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

[Steel, 1994]: The following is a distance function on the edges

$$d_{x,y} = -\ln|\det(P_{x,y})| + \frac{1}{2}\prod_{\sigma \text{ in } \Sigma} \pi_{x,\sigma} - \frac{1}{2}\prod_{\sigma \text{ in } \Sigma} \pi_{y,\sigma}$$

where $P_{x,y}$ is the joint distribution, and the distance between leaves is the sum of distances on the path in the tree

(It's not even obvious it's nonnegative!)

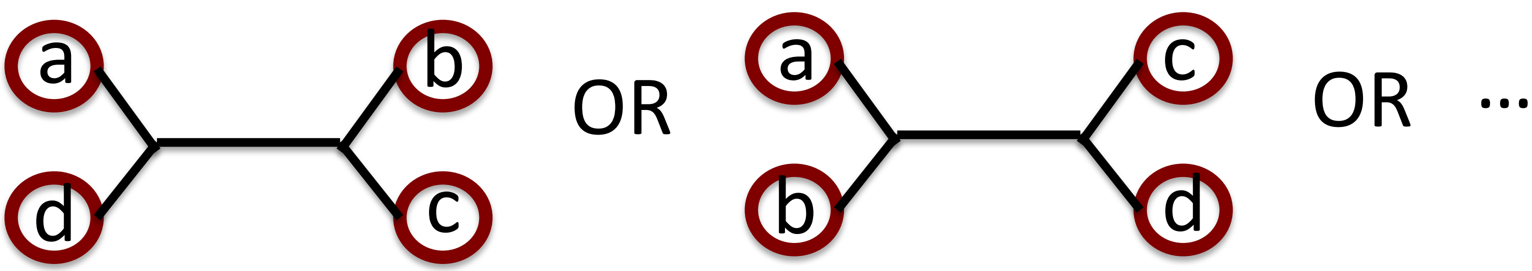
Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

[Erdos, Steel, Szekely, Warnow, 1997]: Used Steel's distance function and quartet tests

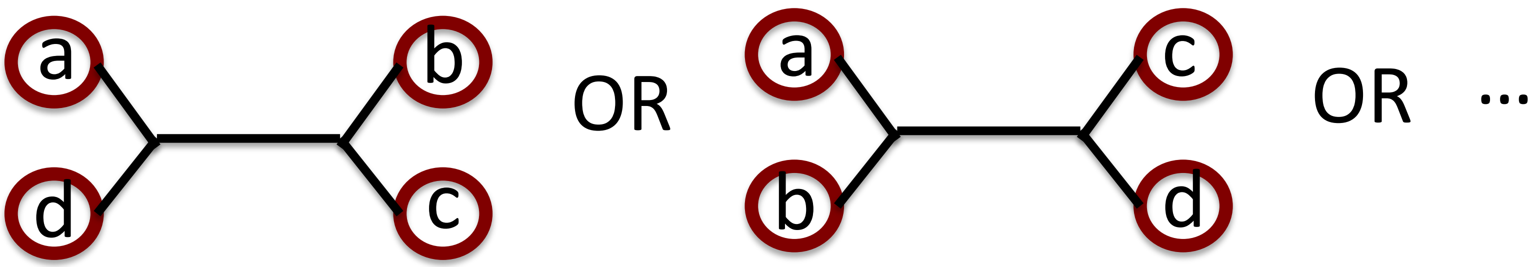


to reconstruction the topology

Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

[Erdos, Steel, Szekely, Warnow, 1997]: Used Steel's distance function and quartet tests

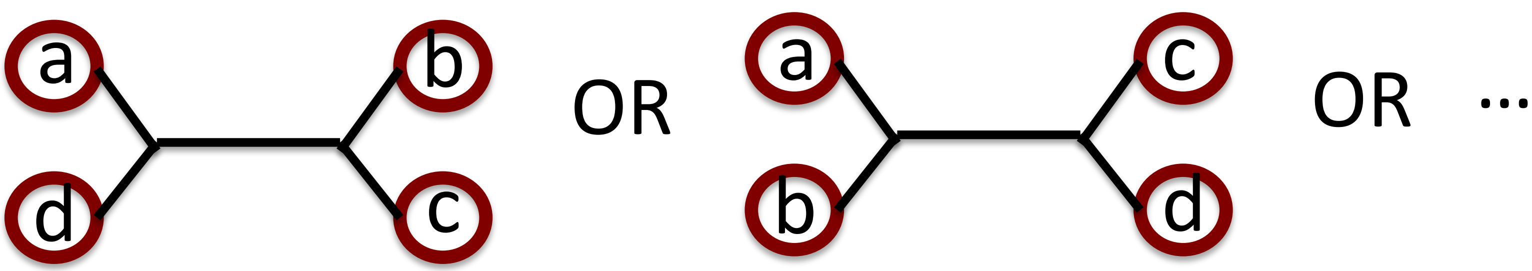


to reconstruction the topology, from polynomially many samples

Can we reconstruct just the topology from random samples?

Usually, we assume $T_{x,y}$, etc are full rank so that we can re-root the tree arbitrarily

[Erdos, Steel, Szekely, Warnow, 1997]: Used Steel's distance function and quartet tests

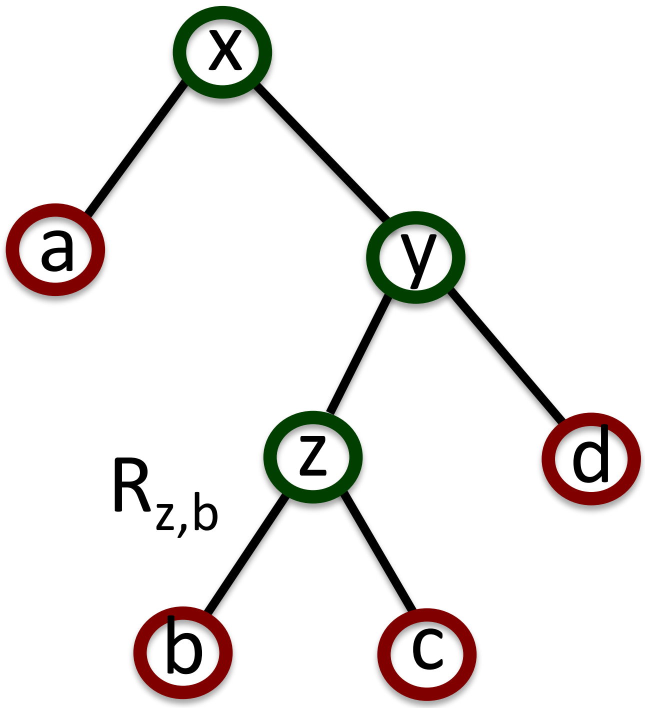


to reconstruction the topology, from polynomially many samples

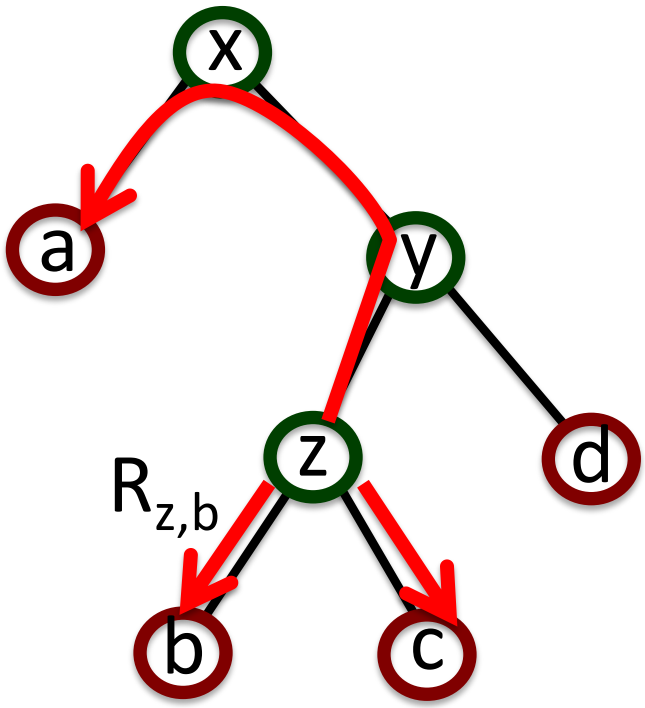
For many problems (e.g. HMMs) finding the transition matrices is the main issue…

[Chang, 1996]: The model is identifiable (if R's are full rank)

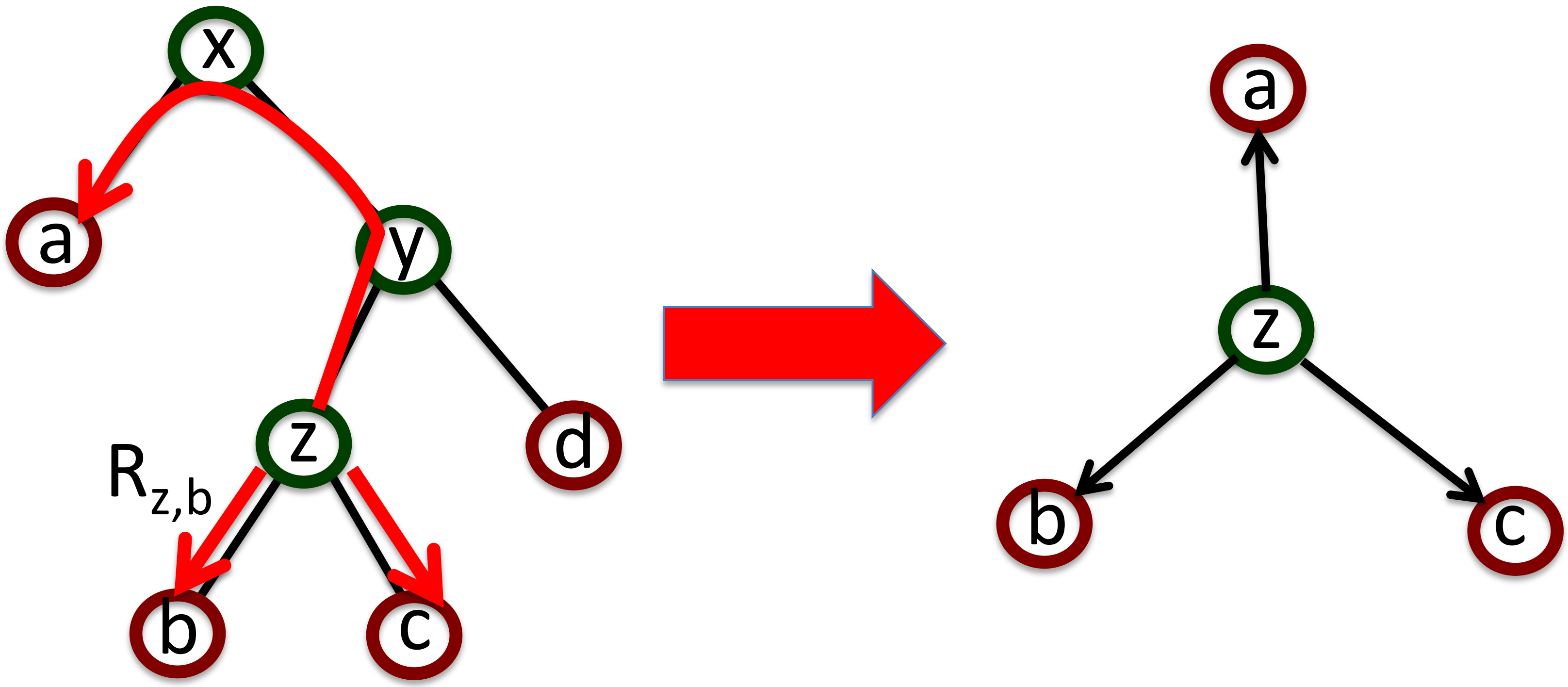
[Chang, 1996]: The model is identifiable (if R's are full rank)



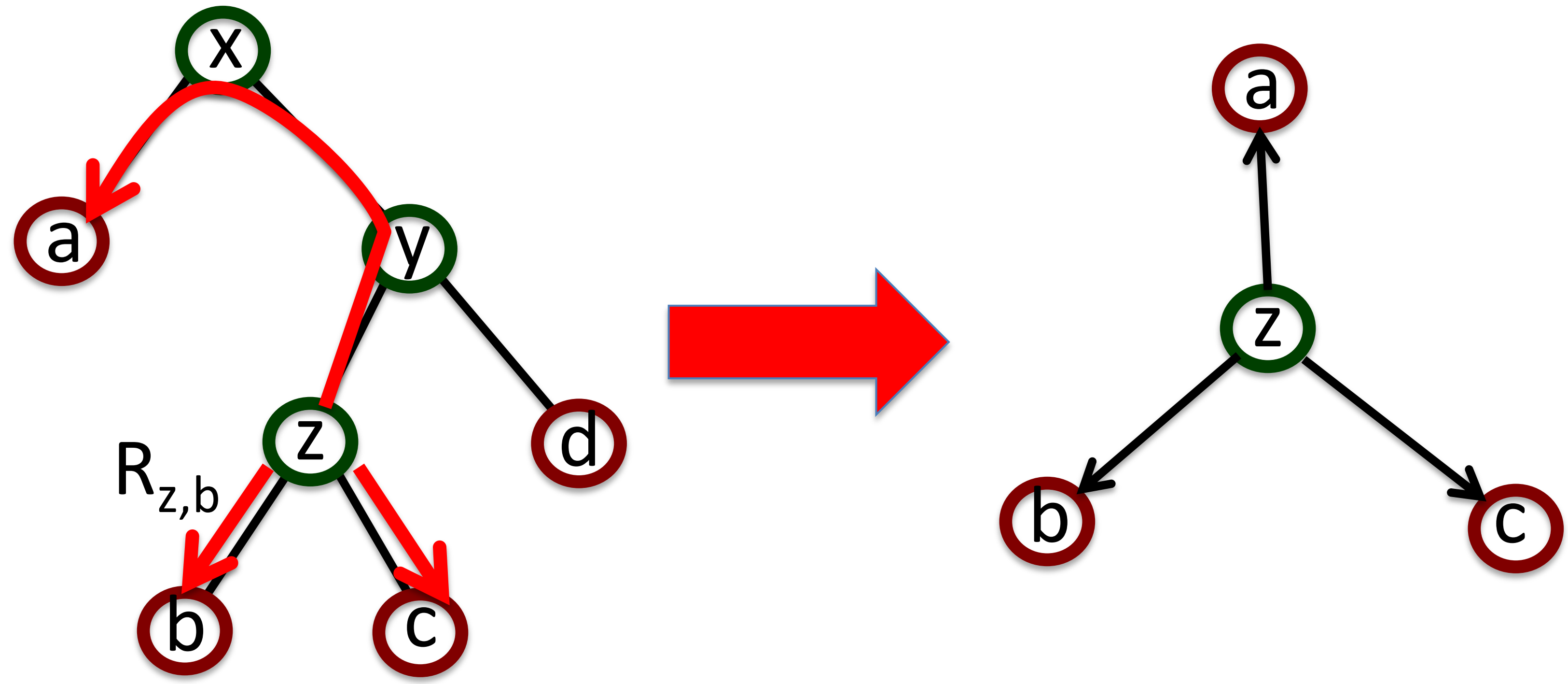
[Chang, 1996]: The model is identifiable (if R's are full rank)



[Chang, 1996]: The model is identifiable (if R's are full rank)



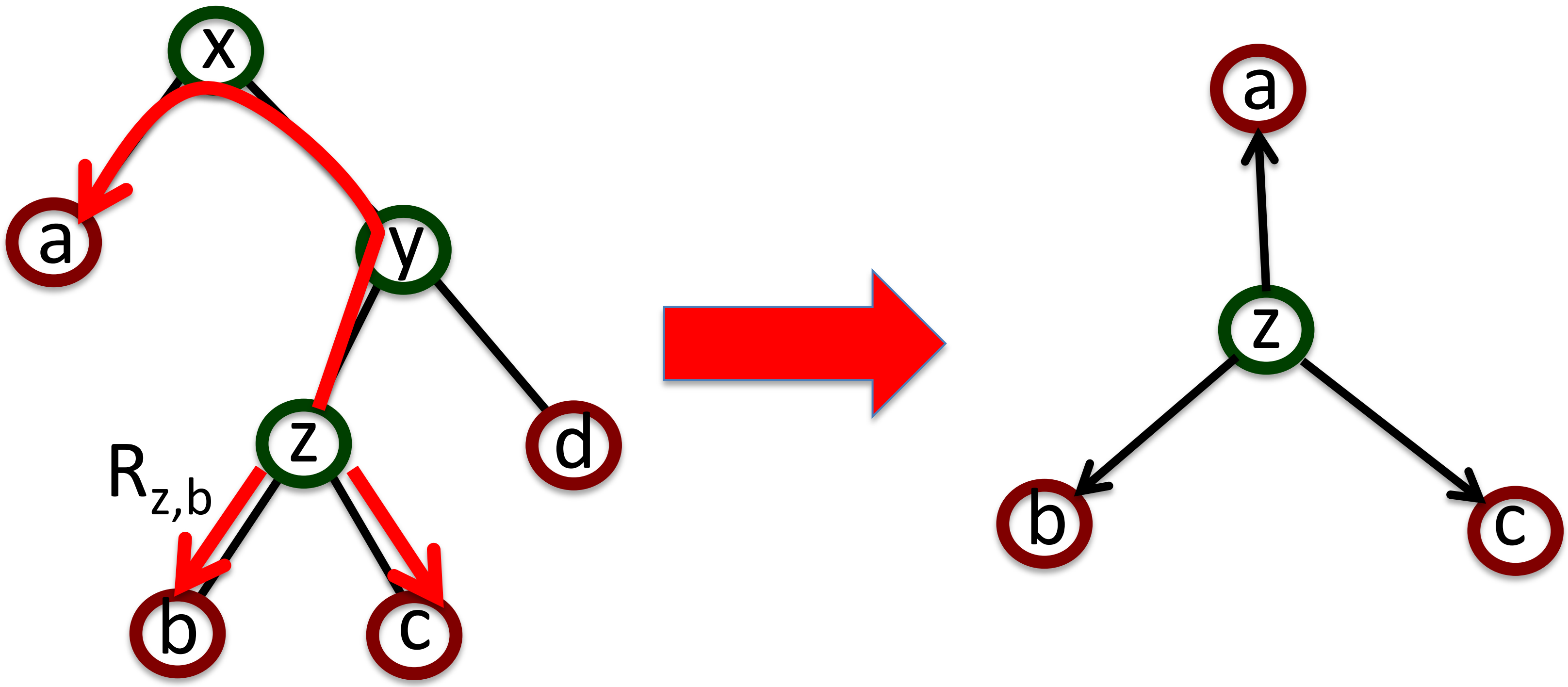
[Chang, 1996]: The model is identifiable (if R's are full rank)



Joint distribution over (a, b, c):

$$\sum_{\sigma} \mathbb{P}[z=\sigma]\mathbb{P}[a|z=\sigma] \otimes \mathbb{P}[b|z=\sigma] \otimes \mathbb{P}[c|z=\sigma]$$

[Chang, 1996]: The model is identifiable (if R's are full rank)



Joint distribution over (a, b, c):

$$\sum_{\sigma} \mathbb{P}[z=\sigma]\mathbb{P}[a|z=\sigma] \otimes \underbrace{\mathbb{P}[b|z=\sigma]}_{\text{columns of } R_{z,b}} \otimes \mathbb{P}[c|z=\sigma]$$

[Mossel, Roch, 2006]: There is an algorithm to PAC learn a phylogenetic tree or an HMM (if its transition/output matrices are full rank) from polynomially many samples

[Mossel, Roch, 2006]: There is an algorithm to PAC learn a phylogenetic tree or an HMM (if its transition/output matrices are full rank) from polynomially many samples

Is the full-rank assumption necessary?

[Mossel, Roch, 2006]: There is an algorithm to PAC learn a phylogenetic tree or an HMM (if its transition/output matrices are full rank) from polynomially many samples

Is the full-rank assumption necessary?

[Mossel, Roch, 2006]: It is as hard as noisy-parity to learn the parameters of a general HMM

[Mossel, Roch, 2006]: There is an algorithm to PAC learn a phylogenetic tree or an HMM (if its transition/output matrices are full rank) from polynomially many samples

Is the full-rank assumption necessary?

[Mossel, Roch, 2006]: It is as hard as noisy-parity to learn the parameters of a general HMM

Noisy-parity is an infamous problem in learning, where O(n) samples suffice but the best algorithms run in time $2^{n/\log(n)}$

Due to **[Blum, Kalai, Wasserman, 2003]**

[Mossel, Roch, 2006]: There is an algorithm to PAC learn a phylogenetic tree or an HMM (if its transition/output matrices are full rank) from polynomially many samples

Is the full-rank assumption necessary?

[Mossel, Roch, 2006]: It is as hard as noisy-parity to learn the parameters of a general HMM

Noisy-parity is an infamous problem in learning, where O(n) samples suffice but the best algorithms run in time $2^{n/\log(n)}$

Due to **[Blum, Kalai, Wasserman, 2003]**

(It's now used as a hard problem to build cryptosystems!)

THE POWER OF CONDITIONAL INDEPENDENCE

[Phylogenetic Trees/HMMS]: (joint distribution on leaves a, b, c)

$$\sum_{\sigma} \mathbb{P}[z=\sigma]\mathbb{P}[a|z=\sigma] \otimes \mathbb{P}[b|z=\sigma] \otimes \mathbb{P}[c|z=\sigma]$$

following **[Mossel, Roch, 2006]**

OUTLINE

Part I: Introduction

- The Rotation Problem
- Jennrich's Algorithm

Part II: Applications

- Phylogenetic Reconstruction
- Mixtures of Gaussians
- Orbit Retrieval

OUTLINE

Part I: Introduction

- The Rotation Problem
- Jennrich's Algorithm

Part II: Applications

- Phylogenetic Reconstruction
- **Mixtures of Gaussians**
- Orbit Retrieval

MIXTURES OF SPHERICAL GAUSSIANS

Let's see another powerful application of tensor methods to learning mixtures of spherical Gaussians

$$\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma^2 I, x)$$

MIXTURES OF SPHERICAL GAUSSIANS

Let's see another powerful application of tensor methods to learning mixtures of spherical Gaussians

$$\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma^2 I, x)$$

Can we reconstruct the parameters in polynomial time?

MIXTURES OF SPHERICAL GAUSSIANS

Let's see another powerful application of tensor methods to learning mixtures of spherical Gaussians

$$\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma^2 I, x)$$

Can we reconstruct the parameters in polynomial time?

Theorem [Hsu, Kakade, 2013]: There is an algorithm that has polynomial run time/sample complexity that works when the μ_i's have full rank

smallest singular value

Running time and sample complexity depend on $1/\sigma_{min}$

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Again, there is a low rank tensor that can be computed from samples whose tensor decomposition reveals the parameters we want to learn

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Case #1: If a, b, c are distinct then we have

$$\mathbb{E}[x_a x_b x_c] = \Big(\sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i \Big)_{a,b,c}$$

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Case #2: If a = b ≠ c then we have

$$\mathbb{E}[x_a x_b x_c] = \Big(\sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i\Big)_{a,b,c} + \sigma^2 \Big(\sum_{i=1}^{k} w_i \mu_i\Big)_c$$

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Case #2: If a = b ≠ c then we have

$$\mathbb{E}[x_a x_b x_c] = \Big(\sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i \Big)_{a,b,c} + \sigma^2 \underbrace{\Big(\sum_{i=1}^{k} w_i \mu_i \Big)_c}_{\text{first moment}}$$

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Case #3: If a = b = c then we have

$$\mathbb{E}[x_a x_b x_c] = \Big(\sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i \Big)_{a,b,c} - 3\sigma^2 \Big(\sum_{i=1}^{k} w_i \mu_i \Big)_c$$

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Proof: Consider the a, b, c entry of the third moment tensor

Case #3: If a = b = c then we have

$$\mathbb{E}[x_a x_b x_c] = \Big(\sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i\Big)_{a,b,c} - 3\sigma^2 \Big(\sum_{i=1}^{k} w_i \mu_i\Big)_c$$

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

It can be written compactly as

$$T = \mathbb{E}[x \otimes x \otimes x] - \sigma^2 \sum_{j=1}^{d} M_j \quad \text{with}$$

$$M_j = \left(\mathbb{E}[x] \otimes e_j \otimes e_j + e_j \otimes \mathbb{E}[x] \otimes e_j + e_j \otimes e_j \otimes \mathbb{E}[x]\right)$$

Main Lemma: If σ^2 is known then the tensor

$$T = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

can be expressed through the empirical moments of the mixture

It can be written compactly as

$$T = \mathbb{E}[x \otimes x \otimes x] - \sigma^2 \sum_{j=1}^{d} M_j \quad \text{with}$$

$$M_j = \left(\mathbb{E}[x] \otimes e_j \otimes e_j + e_j \otimes \mathbb{E}[x] \otimes e_j + e_j \otimes e_j \otimes \mathbb{E}[x]\right)$$

Now use Jennrich's Algorithm

THE POWER OF CONDITIONAL INDEPENDENCE

[Phylogenetic Trees/HMMS]: (joint distribution on leaves a, b, c)

$$\sum_{\sigma} \mathbb{P}[z=\sigma]\mathbb{P}[a|z=\sigma] \otimes \mathbb{P}[b|z=\sigma] \otimes \mathbb{P}[c|z=\sigma]$$

following **[Mossel, Roch, 2006]**

THE POWER OF CONDITIONAL INDEPENDENCE

[Phylogenetic Trees/HMMS]: (joint distribution on leaves a, b, c)

$$\sum_{\sigma} \mathbb{P}[z=\sigma]\mathbb{P}[a|z=\sigma] \otimes \mathbb{P}[b|z=\sigma] \otimes \mathbb{P}[c|z=\sigma]$$

following **[Mossel, Roch, 2006]**

[Mixtures of Spherical Gaussians]: (corrections of third moment)

$$\mathbb{E}[x \otimes x \otimes x] - \sigma^2 \sum_{j=1}^{d} M_j$$

following **[Hsu, Kakade, 2013]**

THE POWER OF CONDITIONAL INDEPENDENCE

[Pure Topic Models/LDA]: (joint distribution on first three words)

$$\sum_j \mathbb{P}[\text{topic} = j] A_j \otimes A_j \otimes A_j$$

following **[Anandkumar, Hsu, Kakade, 2012]**

THE POWER OF CONDITIONAL INDEPENDENCE

[Pure Topic Models/LDA]: (joint distribution on first three words)

$$\sum_j \mathbb{P}[\text{topic} = j] A_j \otimes A_j \otimes A_j$$

following **[Anandkumar, Hsu, Kakade, 2012]**

[Community Detection]: (counting stars)

$$\sum_j \mathbb{P}[C_x = j] \Big(C_A \Pi \Big)_j \otimes \Big(C_B \Pi \Big)_j \otimes \Big(C_C \Pi \Big)_j$$

following **[Anandkumar, Ge, Hsu, Kakade, 2014]**

OUTLINE

Part I: Introduction

- The Rotation Problem
- Jennrich's Algorithm

Part II: Applications

- Phylogenetic Reconstruction
- Mixtures of Gaussians
- Orbit Retrieval

OUTLINE

Part I: Introduction

- The Rotation Problem
- Jennrich's Algorithm

Part II: Applications

- Phylogenetic Reconstruction
- Mixtures of Gaussians
- **Orbit Retrieval**

ORBIT RETRIEVAL

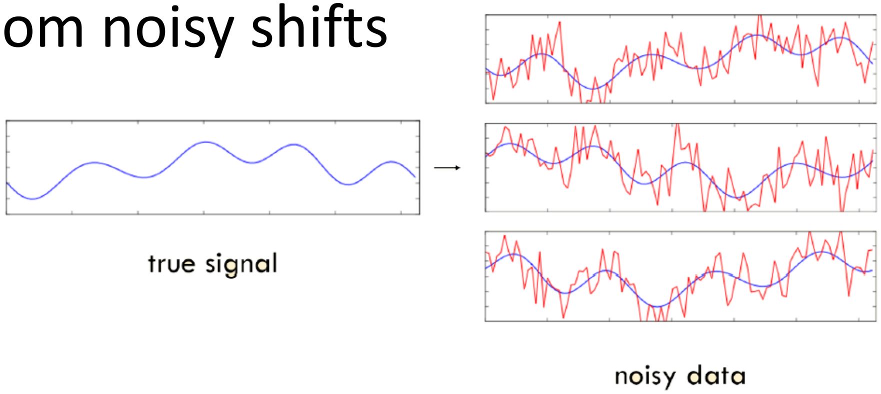
What if we want to learn the parameters of generative model with a continuous latent variable?

ORBIT RETRIEVAL

What if we want to learn the parameters of generative model with a continuous latent variable?

Multireference Alignment

Recover a signal from random noisy shifts



ORBIT RETRIEVAL

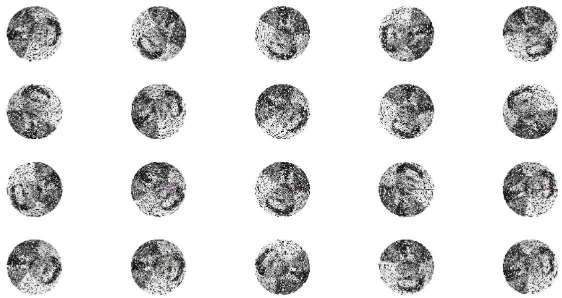
What if we want to learn the parameters of generative model with a continuous latent variable?

ORBIT RETRIEVAL

What if we want to learn the parameters of generative model with a continuous latent variable?

Global Registration

Estimate positions from rigid motions



ORBIT RETRIEVAL

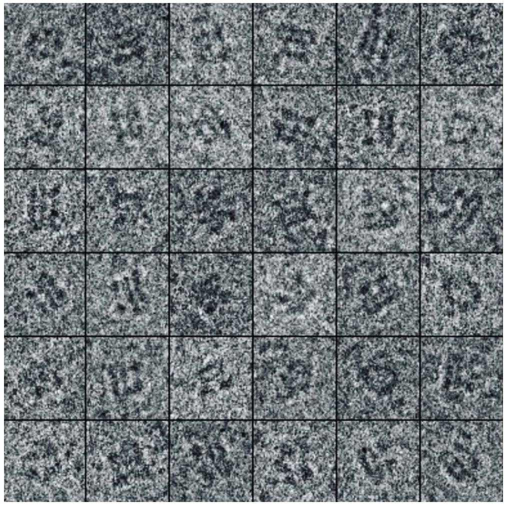
What if we want to learn the parameters of generative model with a continuous latent variable?

ORBIT RETRIEVAL

What if we want to learn the parameters of generative model with a continuous latent variable?

Cryo-electron microscopy

Determine 3D structure from random noisy 2D projections



ORBIT RETRIEVAL

Definition: An **orbit retrieval** problem is specified by a group G and a linear homomorphism

$$\rho : G \to GL(\mathbb{R}^d)$$

We get noisy observations under the group action

$$\rho(g) \cdot x + \eta$$

where g is chosen from the Haar measure on G and η is Gaussian noise

ORBIT RETRIEVAL

Definition: An **orbit retrieval** problem is specified by a group G and a linear homomorphism

$$\rho : G \to GL(\mathbb{R}^d)$$

We get noisy observations under the group action

$$\rho(g) \cdot x + \eta$$

where g is chosen from the Haar measure on G and η is Gaussian noise

Goal: Recover some $\widehat{x}$ that is close to the orbit

$$\{\rho(g) \cdot x | g \in G\}$$

ORBIT TENSOR DECOMPOSITION

In many settings we can estimate

$$T = \int_{g \in G} (\rho(g) \cdot x)^{\otimes 3} dg$$

ORBIT TENSOR DECOMPOSITION

In many settings we can estimate

$$T = \int_{g \in G} (\rho(g) \cdot x)^{\otimes 3} dg$$

Can we recover x up to its orbit?

ORBIT TENSOR DECOMPOSITION

In many settings we can estimate

$$T = \int_{g \in G} (\rho(g) \cdot x)^{\otimes 3} dg$$

Can we recover x up to its orbit?

Theorem [Moitra, Wein, 2019]: There is a polynomial time algorithm that works for SO(2) when x is random

ORBIT TENSOR DECOMPOSITION

In many settings we can estimate

$$T = \int_{g \in G} (\rho(g) \cdot x)^{\otimes 3} dg$$

Can we recover x up to its orbit?

Theorem [Moitra, Wein, 2019]: There is a polynomial time algorithm that works for SO(2) when x is random

What about for non-abelian groups?

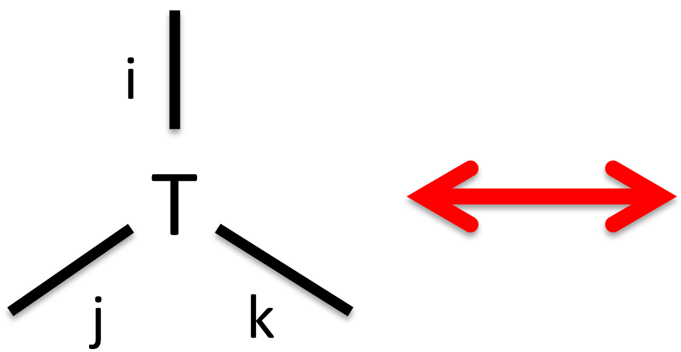
TENSOR NETWORKS

Tensor networks are a graphical representation for tensors and operations on them, e.g.

TENSOR NETWORKS

Tensor networks are a graphical representation for tensors and operations on them, e.g.

third order tensors have three legs

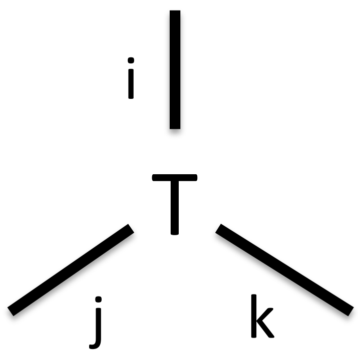


$$T = (T_{i,j,k})$$

TENSOR NETWORKS

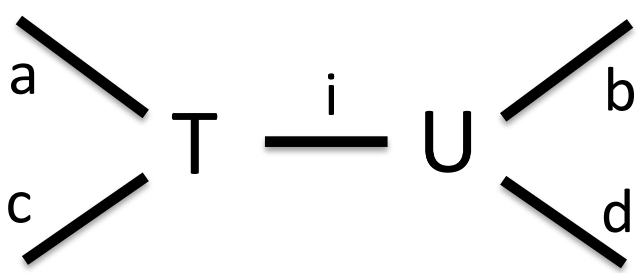
Tensor networks are a graphical representation for tensors and operations on them, e.g.

third order tensors have three legs



$$T = (T_{i,j,k})$$

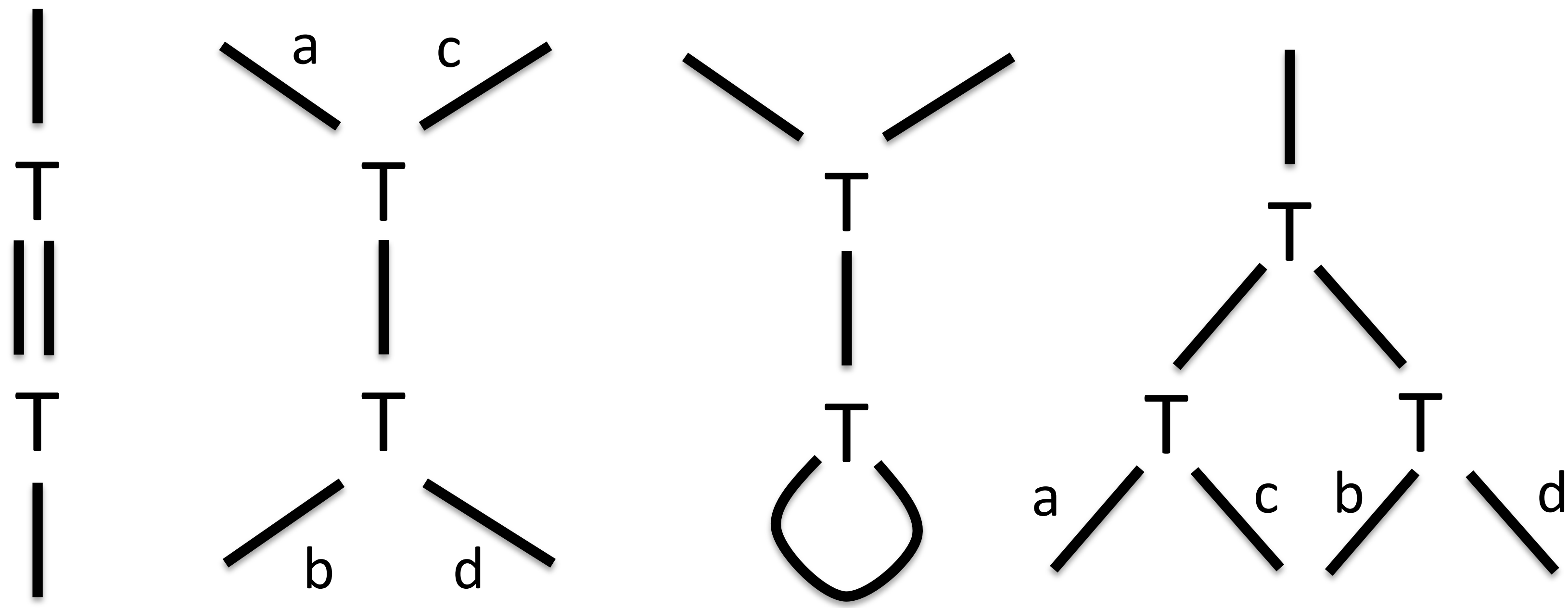
tensors can be attached by summing over connected indices



$$B_{a,b,c,d} = \sum_i T_{a,c,i} U_{b,d,i}$$

REVISITING PRIOR WORK

Prior work implicitly uses this framework



See **[Richard, Montanari]**, **[Barak, Moitra]**, **[Hopkins, Shi, Steurer]**, **[Hopkins et al.]**, **[Hopkins, Shi, Steurer]** for applications to tensor principal component analysis, tensor completion, decomposing random overcomplete third order tensors, etc

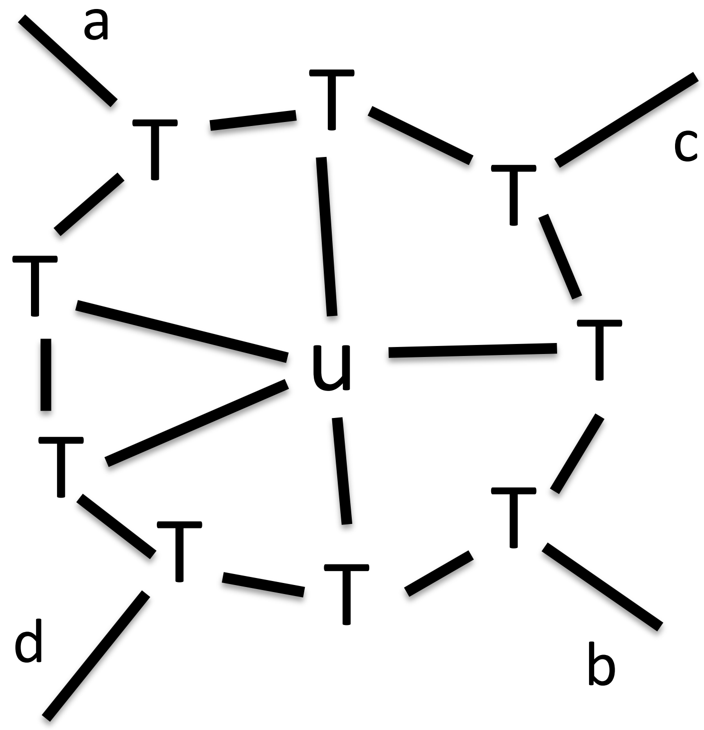
SPECTRAL METHODS FROM TENSOR NETS

Given input tensor T

- **Step #1:** Build a new tensor B by connecting copies of T according to the tensor network
- **Step #2:** Flatten B to form a symmetric matrix M
- **Step #3:** Compute the leading eigenvector of M

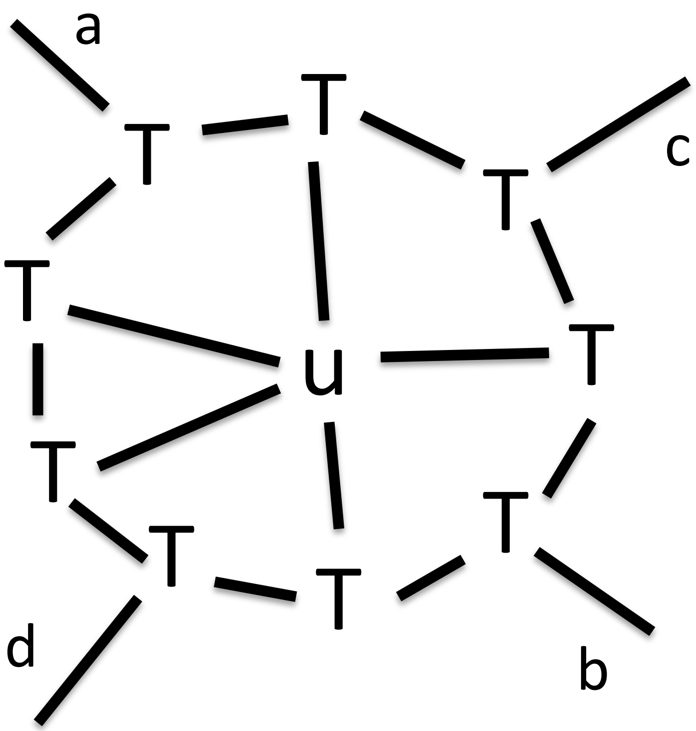
THE BLUEPRINT

We give a spectral method based on the following tensor network



THE BLUEPRINT

We give a spectral method based on the following tensor network



Smaller tensor networks fail for this problem

THE TRACE METHOD

Main step in the analysis is to bound the largest eigenvalue of some matrix build from a tensor network (**after projecting out signal**)

THE TRACE METHOD

Main step in the analysis is to bound the largest eigenvalue of some matrix build from a tensor network (**after projecting out signal**)

We do this through the **trace method**:

$$\mathrm{Tr}(M^{2k}) = \sum_i \lambda_i^{2k} \geq \lambda_{\max}^{2k}$$

THE TRACE METHOD

Main step in the analysis is to bound the largest eigenvalue of some matrix build from a tensor network (**after projecting out signal**)

We do this through the **trace method**:

$$\mathrm{Tr}(M^{2k}) = \sum_i \lambda_i^{2k} \geq \lambda_{\max}^{2k}$$

Applying Markov's inequality we get the bound

$$\mathbb{P}[\lambda_{\max} \geq t] = \mathbb{P}[\lambda_{\max}^{2k} \geq t^{2k}] \leq \frac{\mathbb{E}[\mathrm{Tr}(M^{2k})]}{t^{2k}}$$

THE TRACE METHOD

Main step in the analysis is to bound the largest eigenvalue of some matrix build from a tensor network (**after projecting out signal**)

We do this through the **trace method**:

$$\mathrm{Tr}(M^{2k}) = \sum_i \lambda_i^{2k} \geq \lambda_{\max}^{2k}$$

Applying Markov's inequality we get the bound

$$\mathbb{P}[\lambda_{\max} \geq t] = \mathbb{P}[\lambda_{\max}^{2k} \geq t^{2k}] \leq \frac{\mathbb{E}[\mathrm{Tr}(M^{2k})]}{t^{2k}}$$

With tensor networks, the trace method turns into a counting problem

THE TRACE METHOD

Main step in the analysis is to bound the largest eigenvalue of some matrix build from a tensor network (**after projecting out signal**)

We do this through the **trace method**:

$$\mathrm{Tr}(M^{2k}) = \sum_i \lambda_i^{2k} \geq \lambda_{\max}^{2k}$$

Applying Markov's inequality we get the bound

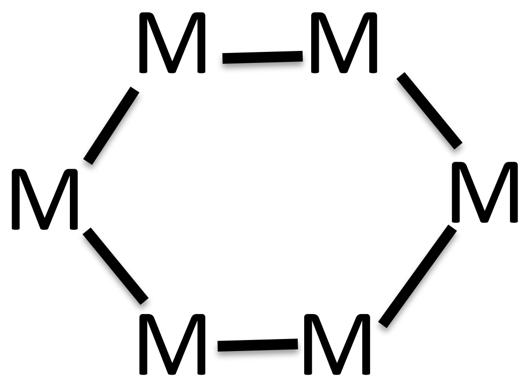
$$\mathbb{P}[\lambda_{\max} \geq t] = \mathbb{P}[\lambda_{\max}^{2k} \geq t^{2k}] \leq \frac{\mathbb{E}[\mathrm{Tr}(M^{2k})]}{t^{2k}}$$

With tensor networks, the trace method turns into a counting problem, **let's see some examples…**

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

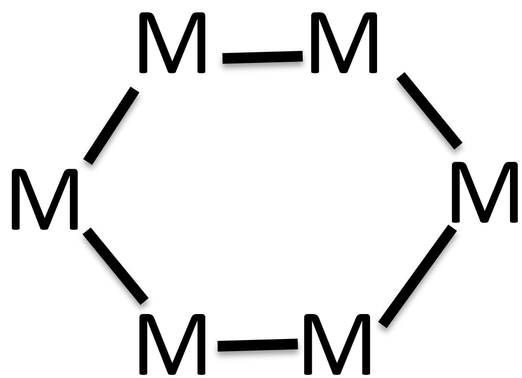
Lemma: $\mathbb{E}[\mathrm{Tr}(M^6)]$ is the number of ways of labeling the edges of



with labels from [n] so that **any pair of labels (i,j) is adjacent to an even number of M's**

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

Lemma: $\mathbb{E}[\mathrm{Tr}(M^6)]$ is the number of ways of labeling the edges of

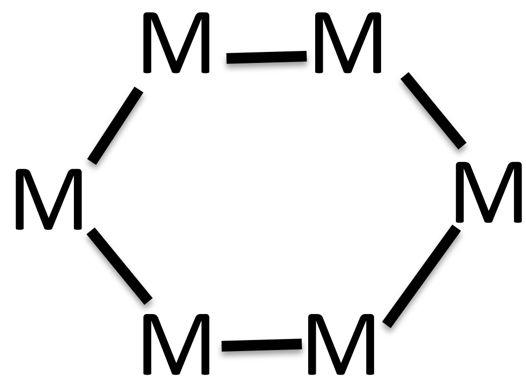


with labels from [n] so that **any pair of labels (i,j) is adjacent to an even number of M's**

Proof: First, $\mathrm{Tr}(M^6)$ is a sum over length six walks.

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

Lemma: $\mathbb{E}[\mathrm{Tr}(M^6)]$ is the number of ways of labeling the edges of



with labels from [n] so that **any pair of labels (i,j) is adjacent to an even number of M's**

Proof: First, $\mathrm{Tr}(M^6)$ is a sum over length six walks. Then observe that a term has expectation zero **unless each edge is traversed an even number of times**.

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

More generally:

Lemma: $\mathbb{E}[\mathrm{Tr}(M^{2k})]$ is the number of ways of labeling the edges of of a length 2k cycle so that any pair (i,j) is adjacent to an even number of M's

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

More generally:

Lemma: $\mathbb{E}[\mathrm{Tr}(M^{2k})]$ is the number of ways of labeling the edges of of a length 2k cycle so that any pair (i,j) is adjacent to an even number of M's

The natural way to double cover edges with a walk is to take the depth first search of a tree

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

More generally:

Lemma: $\mathbb{E}[\mathrm{Tr}(M^{2k})]$ is the number of ways of labeling the edges of of a length 2k cycle so that any pair (i,j) is adjacent to an even number of M's

The natural way to double cover edges with a walk is to take the depth first search of a tree

It turns out this is the dominant contribution:

Theorem [Furedi, Komlos]: $\mathbb{E}[\mathrm{Tr}(M^{2k})] \lesssim n^k$

Suppose M is an n x n symmetric matrix with iid Rademacher entries and zeros along the diagonal

More generally:

Lemma: $\mathbb{E}[\mathrm{Tr}(M^{2k})]$ is the number of ways of labeling the edges of of a length 2k cycle so that any pair (i,j) is adjacent to an even number of M's

The natural way to double cover edges with a walk is to take the depth first search of a tree

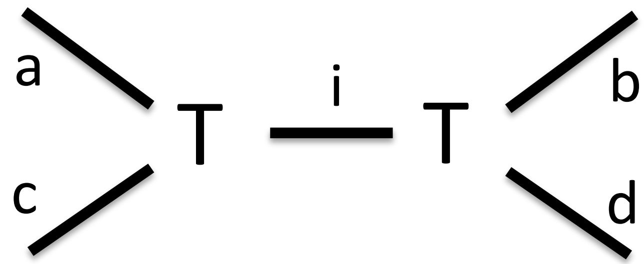
It turns out this is the dominant contribution:

Theorem [Furedi, Komlos]: $\mathbb{E}[\mathrm{Tr}(M^{2k})] \lesssim n^k$

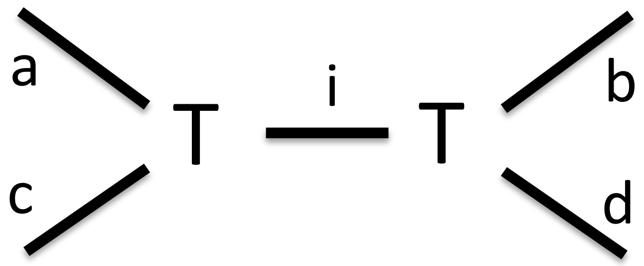
This gives sharp bounds on $||M||$ via the trace method

More challenging example: Suppose T is a symmetric tensor with iid Rademacher entries

More challenging example: Suppose T is a symmetric tensor with iid Rademacher entries and we plug it into the tensor network

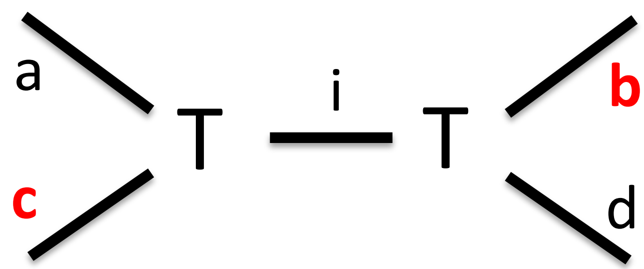


More challenging example: Suppose T is a symmetric tensor with iid Rademacher entries and we plug it into the tensor network



Now let M be the ({a, b}, {c, d})-flattenening

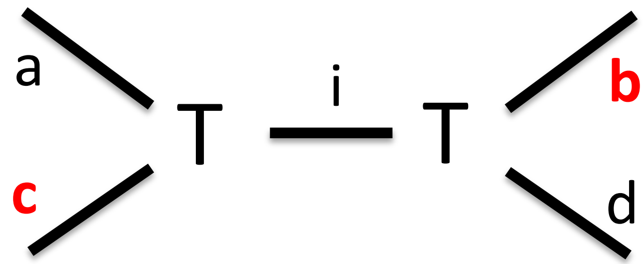
More challenging example: Suppose T is a symmetric tensor with iid Rademacher entries and we plug it into the tensor network



Now let M be the ({a, b}, {c, d})-flattenening

Note that the pair of indices {a,b} that index rows of M come from from different copies of T, and this is important

More challenging example: Suppose T is a symmetric tensor with iid Rademacher entries and we plug it into the tensor network

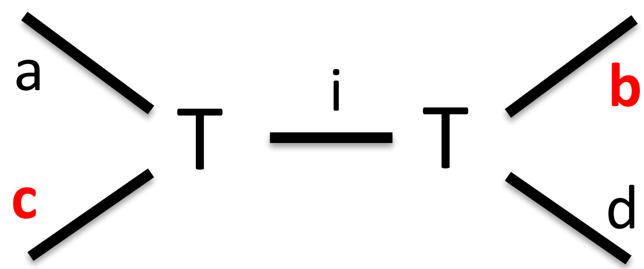


Now let M be the ({a, b}, {c, d})-flattenening

Natural Goal: Understand $||M||$ via the trace method

Note that the pair of indices {a,b} that index rows of M come from from different copies of T, and this is important

More challenging example: Suppose T is a symmetric tensor with iid Rademacher entries and we plug it into the tensor network



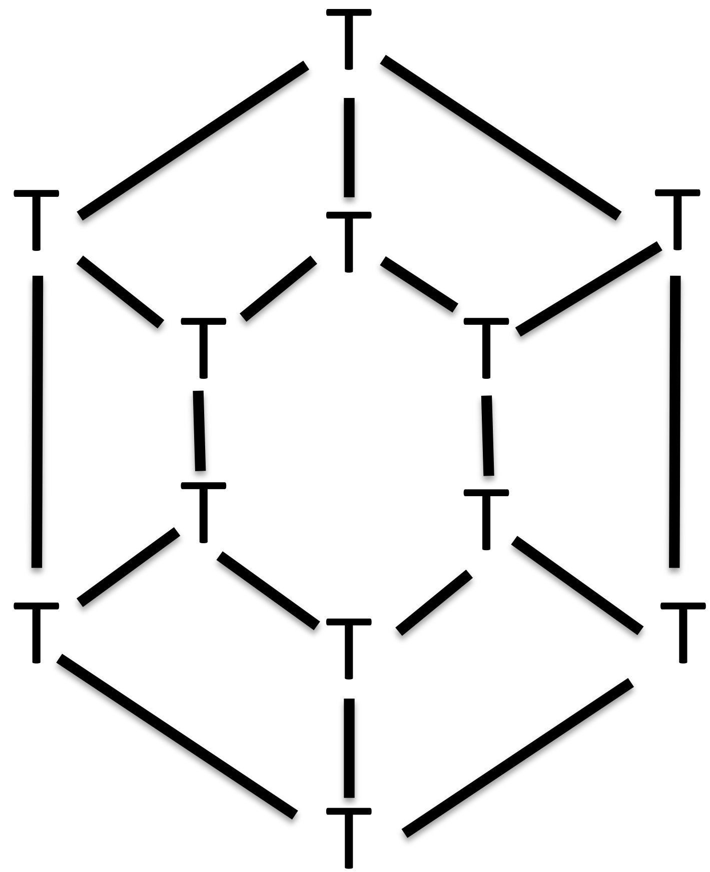
Now let M be the ({a, b}, {c, d})-flattenening

Natural Goal: Understand $||M||$ via the trace method

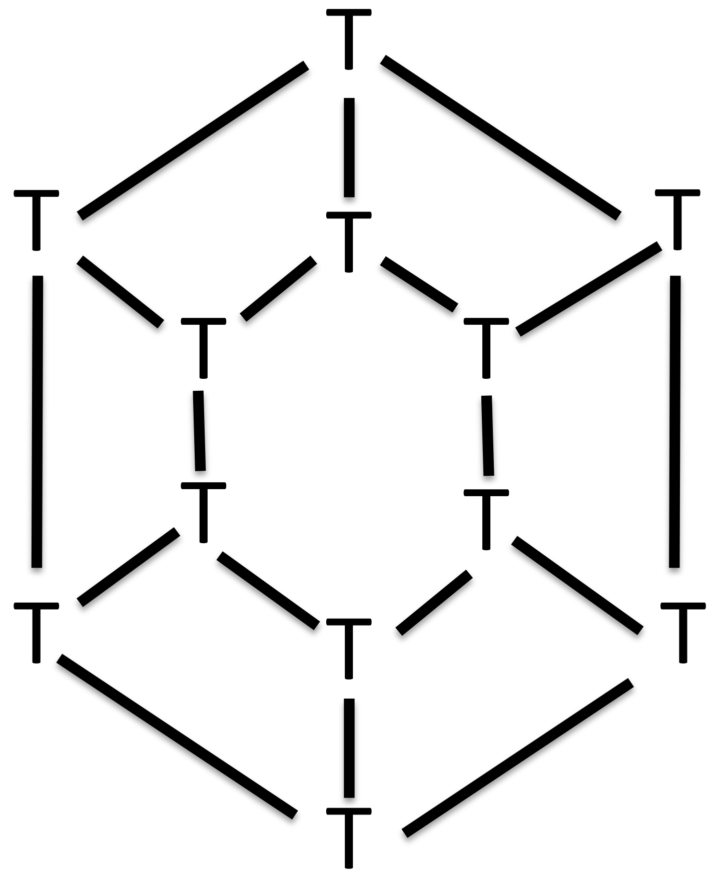
For example, if we want to compute $\mathbb{E}[\mathrm{Tr}(M^6)]$ we can plug the tensor network into the six cycle, and we get…

Note that the pair of indices {a,b} that index rows of M come from from different copies of T, and this is important

…we get:



…we get:



And $\mathbb{E}[\mathrm{Tr}(M^6)]$ is the number of ways to label the edges of the diagram so that **each triple {i, j, k} appears incident to an even number of T's**.

SIDE REMARK

The tensor network formalism gives a visual way to understand some subtleties

SIDE REMARK

The tensor network formalism gives a visual way to understand some subtleties

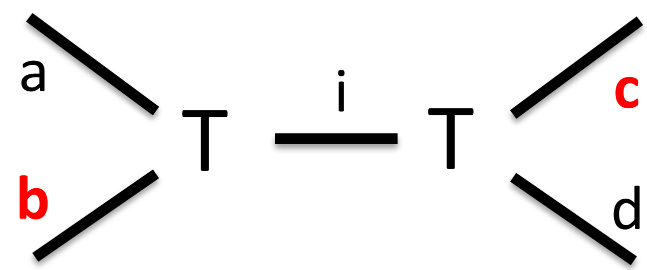
What if we flattened the tensor network differently?

SIDE REMARK

The tensor network formalism gives a visual way to understand some subtleties

What if we flattened the tensor network differently?

For example, if M is the ({a,b}, {c,d})-flattening of

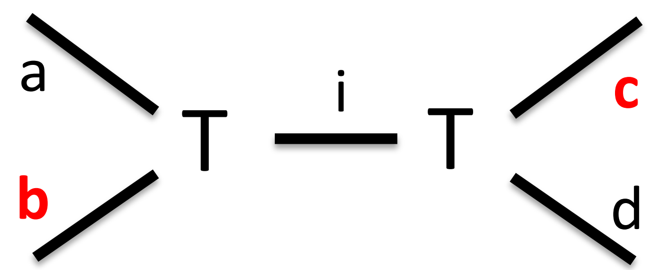


SIDE REMARK

The tensor network formalism gives a visual way to understand some subtleties

What if we flattened the tensor network differently?

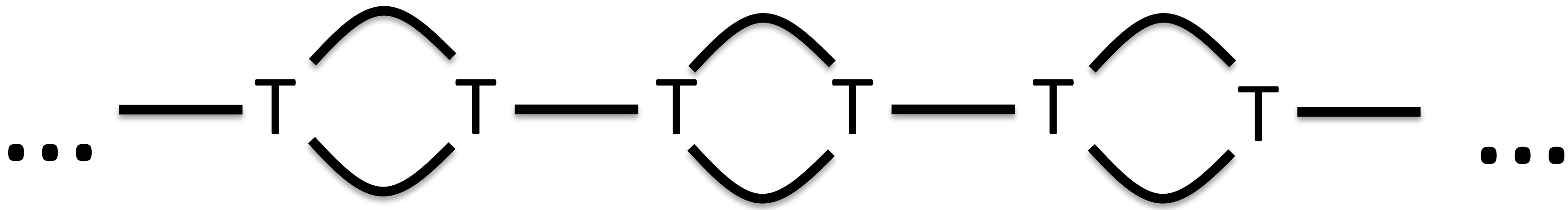
For example, if M is the ({a,b}, {c,d})-flattening of



then plugging it into the six cycle we would get **something different**

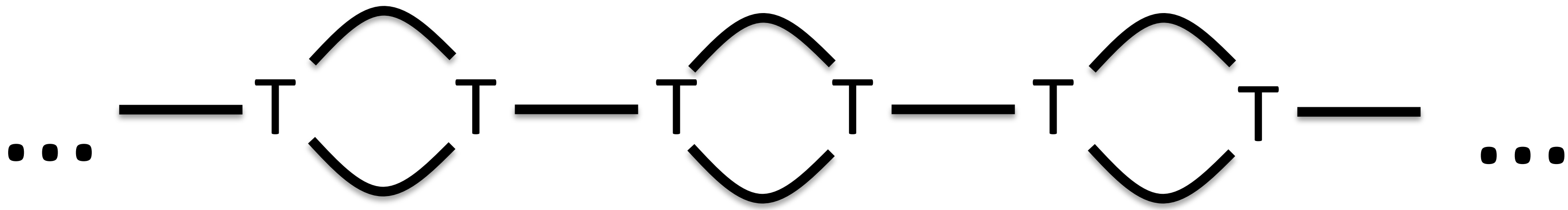
SIDE REMARK

…we get:



SIDE REMARK

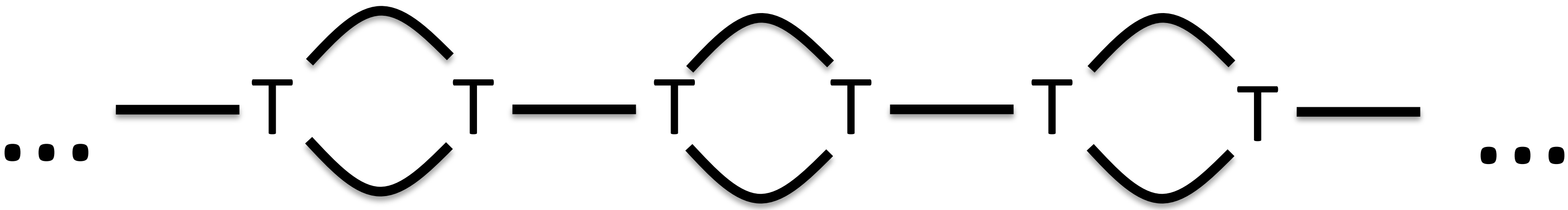
...we get:



Informal Claim: There are now many more labellings where each triple is incident to an even number of T's

SIDE REMARK

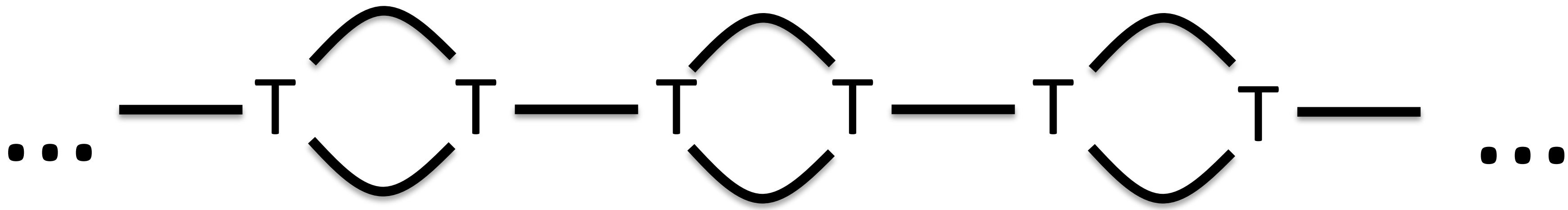
…we get:



Informal Claim: There are now many more labellings where each triple is incident to an even number of T's, **because the graph is only 1-connected**

SIDE REMARK

…we get:



Informal Claim: There are now many more labellings where each triple is incident to an even number of T's, **because the graph is only 1-connected**

Tensor networks are a convenient way to think about this trick, and others that appear in the sum-of-squares literature

TUTORIAL OUTLINE

Part I: Jennrich's Algorithm and its Applications

Part II: Provable Algorithms for Inverse Problems in the Sciences?

Summary:

- Tensor decompositions are unique under more general conditions than matrix decompositions
- Jennrich's Algorithm
- Applications to Phylogenetic Reconstruction, HMMs, Mixtures of Gaussians, Topic Models, …
- **Are there tensor methods that work with group structure?**

Summary:

- Tensor decompositions are unique under more general conditions than matrix decompositions
- Jennrich's Algorithm
- Applications to Phylogenetic Reconstruction, HMMs, Mixtures of Gaussians, Topic Models, …
- **Are there tensor methods that work with group structure?**

Thanks! Any Questions?