

Learning Negative Mixture Models by Tensor Decomposition

Guillaume Rabusseau, François Denis



October 19, 2014

Overview

- 1 Negative Mixtures
- 2 Learning by Tensor Decomposition
- 3 Learning Negative Mixtures
- 4 Experiments
- 5 Conclusion

Overview

1 Negative Mixtures

- Definition
- Negative Mixture of Spherical Gaussians
- Rational Distribution on Strings

2 Learning by Tensor Decomposition

- Tensors: Preliminaries
- Learning from Data
- Tensor Decomposition for Latent Variable Models

3 Learning Negative Mixtures

- Negative Mixture of Spherical Gaussians
- Generalized Tensor Power Method

4 Experiments

5 Conclusion

Mixture Model

Definition

Let f_1, \dots, f_k be distributions and w_1, \dots, w_k be positive reals such that

- $\sum_{i=1}^k w_i = 1$

$f = w_1 f_1 + \dots + w_k f_k$ is a **mixture**.

Mixture Model

Definition

Let f_1, \dots, f_k be distributions and w_1, \dots, w_k be positive reals such that

- $\sum_{i=1}^k w_i = 1$

$f = w_1 f_1 + \dots + w_k f_k$ is a **mixture**.

\mathcal{D}_1 : fair coin, \mathcal{D}_2 : coin s.t. $\mathbb{P}[\text{head}] = 3/4$, $X \sim \mathcal{D} = 0.5\mathcal{D}_1 + 0.5\mathcal{D}_2$

- $\mathbb{P}[X = \text{head}] = 0.5 * 1/2 + 0.5 * 3/4 = 5/8$
- $\mathbb{P}[X = \text{tail}] = 0.5 * 1/2 + 0.5 * 1/4 = 3/8$

Negative Mixture Model

Definition

Let f_1, \dots, f_k be distributions and w_1, \dots, w_k be **non zero** reals such that

- $\sum_{i=1}^k w_i = 1$
- $w_1 f_1(x) + \dots + w_k f_k(x) \geq 0$ for all x

$f = w_1 f_1 + \dots + w_k f_k$ is a **negative (or generalized) mixture**.

Negative Mixture Model

Definition

Let f_1, \dots, f_k be distributions and w_1, \dots, w_k be **non zero** reals such that

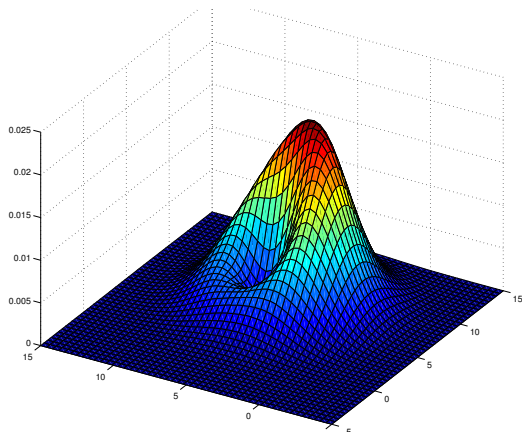
- $\sum_{i=1}^k w_i = 1$
- $w_1 f_1(x) + \dots + w_k f_k(x) \geq 0$ for all x

$f = w_1 f_1 + \dots + w_k f_k$ is a **negative (or generalized) mixture**.

\mathcal{D}_1 : fair coin, \mathcal{D}_2 : coin s.t. $\mathbb{P}[\text{head}] = 3/4$, $X \sim \mathcal{D} = 1.5\mathcal{D}_1 - 0.5\mathcal{D}_2$

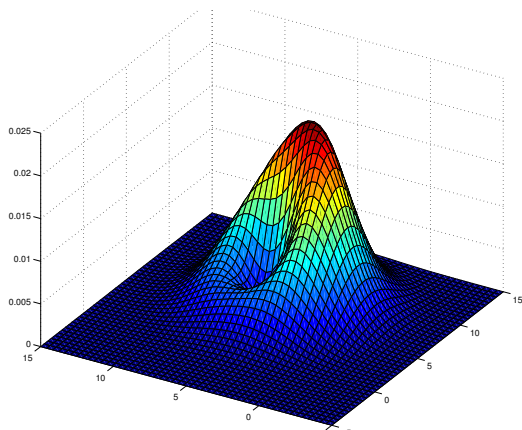
- $\mathbb{P}[X = \text{head}] = 1.5 * 1/2 - 0.5 * 3/4 = 3/8$
- $\mathbb{P}[X = \text{tail}] = 1.5 * 1/2 - 0.5 * 1/4 = 5/8$

Negative Mixture of Spherical Gaussians



$$f(\mathbf{x}) = 1.5\mathcal{N}\left(\mathbf{x}, \begin{bmatrix} 11.4 \\ -3.4 \end{bmatrix}, 8\mathbf{I}\right) - 0.5\mathcal{N}\left(\mathbf{x}, \begin{bmatrix} 11.9 \\ -1.9 \end{bmatrix}, 4\mathbf{I}\right)$$

Negative Mixture of Spherical Gaussians



Simulating $h = \alpha f - (\alpha - 1)g$ by rejection sampling

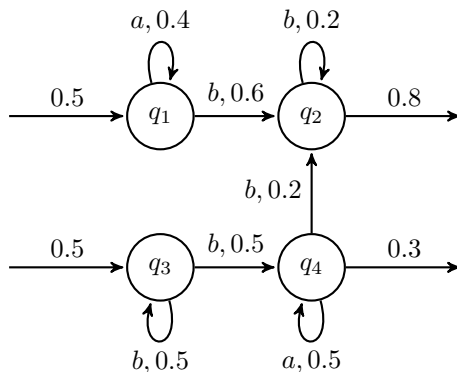
REPEAT

 Draw $x \sim \mathcal{D}_f$ and e uniformly in $[0, 1]$

UNTIL $e\alpha f(x) < (\alpha - 1)g(x)$

RETURN x

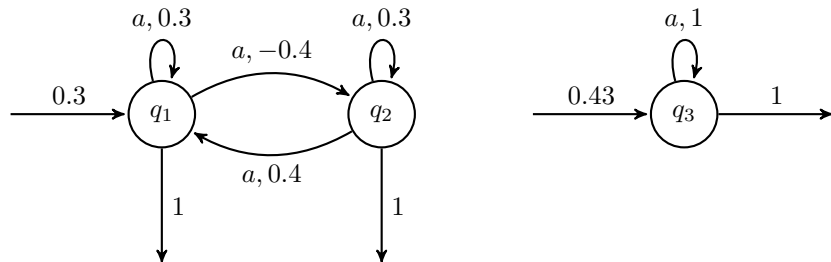
Probabilistic Automata



Example:

$$\mathbb{P}[bb] = 0.5 * 0.6 * 0.2 * 0.8 + 0.5 * 0.5 * 0.5 * 0.3 + 0.5 * 0.5 * 0.2 * 0.8 = 0.1255$$

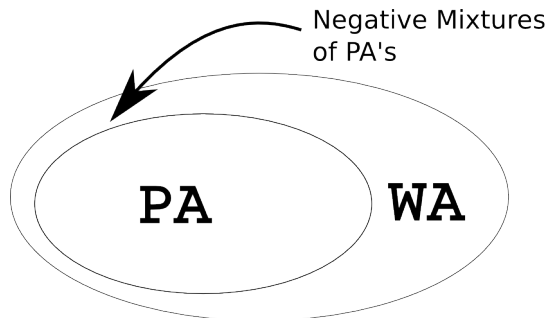
Weighted Automata



Proposition

- $r(w) \geq 0$ for all $w \in \Sigma^*$
- $\sum_{w \in \Sigma^*} r(w) = 1$
- $r \notin PA$

Expressiveness of PA's and WA's



Theorem

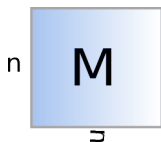
Every rational probability distribution can be generated by the negative mixture of at most two PA's.

Proof: Main difficulty was solved in [Bailly and Denis, 2011].

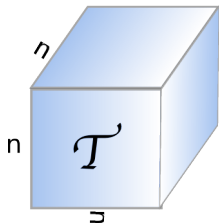
Overview

- 1 Negative Mixtures
 - Definition
 - Negative Mixture of Spherical Gaussians
 - Rational Distribution on Strings
- 2 Learning by Tensor Decomposition
 - Tensors: Preliminaries
 - Learning from Data
 - Tensor Decomposition for Latent Variable Models
- 3 Learning Negative Mixtures
 - Negative Mixture of Spherical Gaussians
 - Generalized Tensor Power Method
- 4 Experiments
- 5 Conclusion

Tensors: Preliminaries

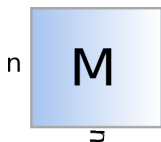


$$\mathbf{M} \in \mathbb{R}^n \otimes \mathbb{R}^n \simeq \mathbb{R}^{n \times n}$$
$$(\mathbf{M}_{ij}) \in \mathbb{R} \text{ for } i, j \in [n]$$

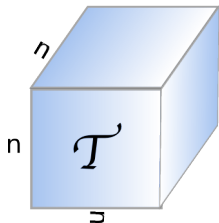


$$\mathcal{T} \in \mathbb{R}^n \otimes \mathbb{R}^n \otimes \mathbb{R}^n$$
$$(\mathcal{T}_{ijk}) \in \mathbb{R} \text{ for } i, j, k \in [n]$$

Tensors: Preliminaries



$$\mathbf{M} \in \mathbb{R}^n \otimes \mathbb{R}^n \simeq \mathbb{R}^{n \times n}$$
$$(\mathbf{M}_{ij}) \in \mathbb{R} \text{ for } i, j \in [n]$$

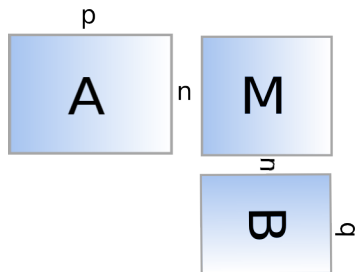


$$\mathcal{T} \in \mathbb{R}^n \otimes \mathbb{R}^n \otimes \mathbb{R}^n$$
$$(\mathcal{T}_{ijk}) \in \mathbb{R} \text{ for } i, j, k \in [n]$$

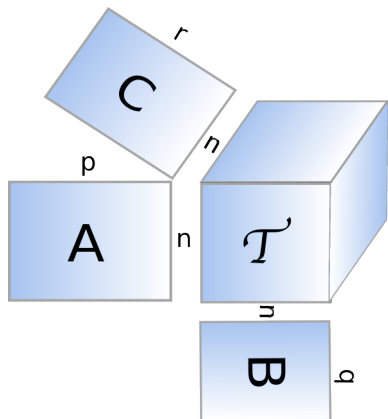
Outer product:

- $\mathbf{a} \otimes \mathbf{b} \in \mathbb{R}^n \otimes \mathbb{R}^n$: $(\mathbf{a} \otimes \mathbf{b})_{ij} = \mathbf{a}_i \mathbf{b}_j$ ($\simeq \mathbf{a} \mathbf{b}^T$)
- $\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} \in \mathbb{R}^n \otimes \mathbb{R}^n \otimes \mathbb{R}^n$: $(\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c})_{ijk} = \mathbf{a}_i \mathbf{b}_j \mathbf{c}_k$

Tensors: Preliminaries

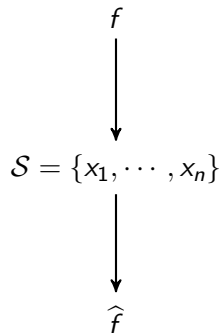


$$\mathbf{A} \in \mathbb{R}^{n \times p}, \mathbf{B} \in \mathbb{R}^{n \times q}$$
$$\mathbf{M}(\mathbf{A}, \mathbf{B}) = \mathbf{A}^T \mathbf{M} \mathbf{B} \in \mathbb{R}^{p \times q}$$



$$\mathbf{A} \in \mathbb{R}^{n \times p}, \mathbf{B} \in \mathbb{R}^{n \times q}, \mathbf{C} \in \mathbb{R}^{n \times r}$$
$$\mathcal{T}(\mathbf{A}, \mathbf{B}, \mathbf{C}) \in \mathbb{R}^p \otimes \mathbb{R}^q \otimes \mathbb{R}^r$$

Learning from Data



Learning from Data: Gaussian

$$\mathcal{N}(x; \mu, \sigma^2)$$



$$\mathcal{S} = \{x_1, \dots, x_n\}$$



$$\begin{cases} \mathbb{E}[x] &= \mu & \simeq \frac{1}{n} \sum_{i=1}^n x_i \\ \mathbb{E}[x^2] &= \sigma^2 + \mu^2 & \simeq \frac{1}{n} \sum_{i=1}^n x_i^2 \end{cases}$$



$$\hat{\mu}, \hat{\sigma}^2$$

Learning from Data: Method of Moments

$$f(x; \theta_1, \dots, \theta_k)$$



$$\mathcal{S} = \{x_1, \dots, x_n\}$$



$$\left\{ \begin{array}{l} \mathbb{E}[x] = g_1(\theta_1, \dots, \theta_k) \simeq \frac{1}{n} \sum_{i=1}^n x_i \\ \mathbb{E}[x^2] = g_2(\theta_1, \dots, \theta_k) \simeq \frac{1}{n} \sum_{i=1}^n x_i^2 \\ \vdots \\ \mathbb{E}[x^k] = g_k(\theta_1, \dots, \theta_k) \simeq \frac{1}{n} \sum_{i=1}^n x_i^k \end{array} \right.$$



$$\hat{\theta}_1, \dots, \hat{\theta}_k$$

Tensor Decomposition for Learning Latent Variable Models

[Anandkumar et al., 2012]

Latent Variable Model:

$$f(\mathbf{x}) = \sum_{i=1}^k p_i f_i(\mathbf{x}; \boldsymbol{\mu}_i)$$



$$\mathcal{S} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$$



Structure in the
Low Order Moments

$$\begin{cases} \mathbb{E}[\mathbf{x} \otimes \mathbf{x}] & = g_1(\sum_{i=1}^k p_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i) \\ \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}] & = g_2(\sum_{i=1}^k p_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i) \end{cases}$$



Tensor Power Method

$$\hat{p}_i, \hat{\boldsymbol{\mu}}_i$$

Mixture of Spherical Gaussians

- Generative process:
 - ▶ Draw a gaussian $h \sim \mathbf{p} \in \mathbb{R}^k$
 - ▶ Draw $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_h, \sigma_h^2 \mathbf{I})$
- The PDF of \mathbf{x} is $p_1 \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}) + \dots + p_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \sigma_k^2 \mathbf{I})$

Theorem ([Hsu and Kakade, 2013], part 1)

The average variance $\bar{\sigma}^2 = \sum_{i=1}^k p_i \sigma_i^2$ is the smallest eigenvalue of the covariance matrix $\mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top]$.

Mixture of Spherical Gaussians

- Generative process:
 - ▶ Draw a gaussian $h \sim \mathbf{p} \in \mathbb{R}^k$
 - ▶ Draw $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_h, \sigma_h^2 \mathbf{I})$
- The PDF of \mathbf{x} is $p_1 \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}) + \dots + p_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \sigma_k^2 \mathbf{I})$

Theorem ([Hsu and Kakade, 2013], part 2)

Let \mathbf{v} be any unit-norm eigenvector corresponding to $\bar{\sigma}^2$ and let

$$\mathbf{m}_1 = \mathbb{E}[\mathbf{x}(\mathbf{v}^\top (\mathbf{x} - \mathbb{E}[\mathbf{x}]))^2], \quad \mathbf{M}_2 = \mathbb{E}[\mathbf{x} \otimes \mathbf{x}] - \bar{\sigma}^2 \mathbf{I}, \quad \text{and}$$
$$\mathcal{M}_3 = \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}] - \sum_{i=1}^n [\mathbf{m}_1 \otimes \mathbf{e}_i \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \mathbf{m}_1 \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{m}_1]$$

where $\mathbf{e}_1, \dots, \mathbf{e}_n$ is the coordinate basis of \mathbb{R}^n . Then,

$$\mathbf{m}_1 = \sum_{i=1}^k p_i \sigma_i^2 \boldsymbol{\mu}_i, \quad \mathbf{M}_2 = \sum_{i=1}^k p_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i, \quad \text{and} \quad \mathcal{M}_3 = \sum_{i=1}^k p_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i.$$

Tensor Power Method: Problem Formulation

$$\begin{cases} \widehat{\mathbf{M}}_2 & \simeq \sum_{i=1}^k p_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \\ \widehat{\mathbf{M}}_3 & \simeq \sum_{i=1}^k p_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \end{cases}$$

↓ ?

$$\widehat{p}_i, \widehat{\boldsymbol{\mu}}_i$$

- $k \leq d$
- $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ are linearly independent
- $p_1, \dots, p_k \in \mathbb{R}$ are **strictly positive real numbers**

Tensor Power Method: Orthonormalization

- $\mathbf{M}_2 = \sum_{i=1}^k w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ eigendecomposition of \mathbf{M}_2 .
- $\mathbf{W} = \mathbf{U}\mathbf{D}^{-1/2} \in \mathbb{R}^{d \times k}$ and $\widetilde{\boldsymbol{\mu}}_i = \sqrt{w_i} \mathbf{W}^\top \boldsymbol{\mu}_i \in \mathbb{R}^k$.
- We have

$$\mathbf{I} = \mathbf{M}_2(\mathbf{W}, \mathbf{W}) = \mathbf{W}^\top \left(\sum_{i=1}^k w_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \right) \mathbf{W} = \sum_{i=1}^k \widetilde{\boldsymbol{\mu}}_i \widetilde{\boldsymbol{\mu}}_i^\top$$

hence $\widetilde{\boldsymbol{\mu}}_i^\top \widetilde{\boldsymbol{\mu}}_j = \delta_{ij}$ for all i, j .

- $\widetilde{\mathcal{M}}_3 = \mathcal{M}_3(\mathbf{W}, \mathbf{W}, \mathbf{W}) = \sum_{i=1}^k \frac{1}{w_i} \widetilde{\boldsymbol{\mu}}_i \otimes \widetilde{\boldsymbol{\mu}}_i \otimes \widetilde{\boldsymbol{\mu}}_i$.

Tensor Power Method

Theorem ([Anandkumar et al., 2012])

Let $\mathcal{T} \in \otimes^3 \mathbb{R}^d$ have an orthonormal decomposition

$$\mathcal{T} = \sum_{i=1}^k \frac{1}{w_i} \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i.$$

Let $\theta_0 \in \mathbb{R}^d$, suppose that $|\frac{1}{w_1} \tilde{\mu}_1^\top \theta_0| > |\frac{1}{w_2} \tilde{\mu}_2^\top \theta_0| \geq \dots \geq |\frac{1}{w_k} \tilde{\mu}_k^\top \theta_0| > 0$.
For $t = 1, 2, \dots$, define

$$\theta_t = \frac{\mathcal{T}(I, \theta_{t-1}, \theta_{t-1})}{\|\mathcal{T}(I, \theta_{t-1}, \theta_{t-1})\|} \quad \text{and} \quad \lambda_t = \mathcal{T}(\theta_t, \theta_t, \theta_t)$$

Then, $\theta_t \rightarrow \tilde{\mu}_1$ and $\lambda_t \rightarrow \frac{1}{w_1}$.

Overview

- 1 Negative Mixtures
 - Definition
 - Negative Mixture of Spherical Gaussians
 - Rational Distribution on Strings
- 2 Learning by Tensor Decomposition
 - Tensors: Preliminaries
 - Learning from Data
 - Tensor Decomposition for Latent Variable Models
- 3 Learning Negative Mixtures
 - Negative Mixture of Spherical Gaussians
 - Generalized Tensor Power Method
- 4 Experiments
- 5 Conclusion

Learning Negative Mixtures by Tensor Decomposition

Negative Mixture Model:

$$f(x) = \sum_{i=1}^k w_i f_i(x; \mu_i)$$



$$\mathcal{S} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$$



Structure in the
Low Order Moments

$$\begin{cases} \mathbb{E}[\mathbf{x} \otimes \mathbf{x}] & = g_1(\sum_{i=1}^k w_i \mu_i \otimes \mu_i) \\ \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}] & = g_2(\sum_{i=1}^k w_i \mu_i \otimes \mu_i \otimes \mu_i) \end{cases}$$



Generalized Tensor
Power Method

$$\hat{w}_i, \hat{\mu}_i$$

Negative Mixture of Spherical Gaussians

$$f(\mathbf{x}) = \sum_{i=1}^k w_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \sigma_i^2 \mathbf{I}), \quad w_i \neq 0$$

Let ℓ be the number of negative coefficients.

Theorem (part 1)

- The average variance $\bar{\sigma}^2 = \sum_{i=1}^k w_i \sigma_i^2$ is an eigenvalue of the covariance matrix $\mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top]$.
- Furthermore, $\bar{\sigma}^2$ is either the ℓ -th or $(\ell + 1)$ -th smallest eigenvalue of the covariance matrix.

Negative Mixture of Spherical Gaussians

$$f(\mathbf{x}) = \sum_{i=1}^k w_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \sigma_i^2 \mathbf{I}), \quad w_i \neq 0$$

Let ℓ be the number of negative coefficients.

Theorem (part 2)

Let \mathbf{v} be any unit-norm eigenvector corresponding to $\bar{\sigma}^2$ and let

$$\mathbf{m}_1 = \mathbb{E}[\mathbf{x}(\mathbf{v}^\top (\mathbf{x} - \mathbb{E}[\mathbf{x}]))^2], \quad \mathbf{M}_2 = \mathbb{E}[\mathbf{x} \otimes \mathbf{x}] - \bar{\sigma}^2 \mathbf{I}, \quad \text{and}$$
$$\mathcal{M}_3 = \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}] - \sum_{i=1}^n [\mathbf{m}_1 \otimes \mathbf{e}_i \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \mathbf{m}_1 \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{m}_1]$$

where $\mathbf{e}_1, \dots, \mathbf{e}_n$ is the coordinate basis of \mathbb{R}^n . Then,

$$\mathbf{m}_1 = \sum_{i=1}^k w_i \sigma_i^2 \boldsymbol{\mu}_i, \quad \mathbf{M}_2 = \sum_{i=1}^k w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i, \quad \text{and} \quad \mathcal{M}_3 = \sum_{i=1}^k w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i.$$

Generalized Tensor Power Method: Problem Formulation

$$\begin{cases} \widehat{\mathbf{M}}_2 & \simeq \sum_{i=1}^k w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \\ \widehat{\mathbf{M}}_3 & \simeq \sum_{i=1}^k w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \end{cases}$$

\downarrow ?

$\widehat{w}_i, \widehat{\boldsymbol{\mu}}_i$

- $k \leq d$
- $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ are linearly independent
- $w_1, \dots, w_k \in \mathbb{R}$ are **non zero**

Generalized Tensor Power Method: Pseudo-Orthonormalization

- $\mathbf{M}_2 = \sum_{i=1}^k w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ eigendecomposition of \mathbf{M}_2 .
- $\mathbf{W} = \mathbf{U}\mathbf{D}^{-1/2} \in \mathbb{C}^{d \times k}$ and $\widetilde{\boldsymbol{\mu}}_i = \sqrt{w_i} \mathbf{W}^\top \boldsymbol{\mu}_i \in \mathbb{C}^k$.
- We have

$$\mathbf{I} = \mathbf{M}_2(\mathbf{W}, \mathbf{W}) = \mathbf{W}^\top \left(\sum_{i=1}^k w_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \right) \mathbf{W} = \sum_{i=1}^k \widetilde{\boldsymbol{\mu}}_i \widetilde{\boldsymbol{\mu}}_i^\top$$

hence $\widetilde{\boldsymbol{\mu}}_i^\top \widetilde{\boldsymbol{\mu}}_j = \delta_{ij}$ for all i, j .

- $\widetilde{\mathcal{M}}_3 = \mathcal{M}_3(\mathbf{W}, \mathbf{W}, \mathbf{W}) = \sum_{i=1}^k \frac{1}{w_i} \widetilde{\boldsymbol{\mu}}_i \otimes \widetilde{\boldsymbol{\mu}}_i \otimes \widetilde{\boldsymbol{\mu}}_i \quad (\in \mathbb{C}^k \otimes \mathbb{C}^k \otimes \mathbb{C}^k)$

(!) $\langle \widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\mu}}_j \rangle \neq \widetilde{\boldsymbol{\mu}}_i^\top \widetilde{\boldsymbol{\mu}}_j \in \mathbb{C}$

Example: if $\mathbf{x} = (1 \ i)^\top$ and $\mathbf{y} = (1+i \ 1)^\top$ then $\mathbf{x}^\top \mathbf{x} = 0$ and $\mathbf{y}^\top \mathbf{y} = 1 + 2i$.

Generalized Tensor Power Method: Complex Tensors

Theorem

Let $\mathcal{T} \in \bigotimes^3 \mathbb{C}^n$ have a pseudo-orthonormal decomposition

$$\mathcal{T} = \sum_{i=1}^k \frac{1}{w_i} \tilde{\boldsymbol{\mu}}_i \otimes \tilde{\boldsymbol{\mu}}_i \otimes \tilde{\boldsymbol{\mu}}_i.$$

Let $\boldsymbol{\theta}_0 \in \mathbb{C}^n$, suppose that $|\frac{1}{w_1} \tilde{\boldsymbol{\mu}}_1^\top \boldsymbol{\theta}_0| > |\frac{1}{w_2} \tilde{\boldsymbol{\mu}}_2^\top \boldsymbol{\theta}_0| \geq \dots \geq |\frac{1}{w_k} \tilde{\boldsymbol{\mu}}_k^\top \boldsymbol{\theta}_0| > 0$.
For $t = 1, 2, \dots$, define

$$\boldsymbol{\theta}_t = \frac{\mathcal{T}(I, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-1})}{[\mathcal{T}(I, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-1})^\top \mathcal{T}(I, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-1})]^{1/2}} \quad \text{and} \quad \lambda_t = \mathcal{T}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t, \boldsymbol{\theta}_t)$$

Then, $\boldsymbol{\theta}_t \rightarrow \pm \tilde{\boldsymbol{\mu}}_1$ and $\lambda_t \rightarrow \pm \frac{1}{w_1}$.

Generalized Tensor Power Method: Overall Procedure

Algorithm 1 Negative Mixture Estimation

Input: $k \in \mathbb{N}$, $\widehat{\mathbf{M}}_2 \in \otimes^2 \mathbb{R}^n$, $\widehat{\mathcal{M}}_3 \in \otimes^3 \mathbb{R}^n$

Output: $w_1, \dots, w_k, \mu_1, \dots, \mu_k$

$\mathbf{U}\mathbf{D}\mathbf{U}^\top \leftarrow \widehat{\mathbf{M}}_2$ (k -truncated eig. decomp.);

$\mathbf{W} \leftarrow \mathbf{U}\mathbf{D}^{-\frac{1}{2}}$; $\mathcal{J} \leftarrow \widehat{\mathcal{M}}_3(\mathbf{W}, \mathbf{W}, \mathbf{W})$;

for $i = 1$ **to** k **do**

 Draw θ at random in \mathbb{C}^k ;

repeat

$$\theta \leftarrow \mathcal{J}(I, \theta, \theta); \theta \leftarrow \frac{\theta}{(\theta^\top \theta)^{\frac{1}{2}}};$$

until stabilization

$\lambda \leftarrow \mathcal{J}(\theta, \theta, \theta)$; $\mathcal{J} \leftarrow \mathcal{J} - \lambda \cdot \theta \otimes \theta \otimes \theta$;

$w_i \leftarrow 1/\lambda^2$; $\mu_i \leftarrow \lambda(\mathbf{W}^\top)^+ \theta$;

end for

Overview

- 1 Negative Mixtures
 - Definition
 - Negative Mixture of Spherical Gaussians
 - Rational Distribution on Strings
- 2 Learning by Tensor Decomposition
 - Tensors: Preliminaries
 - Learning from Data
 - Tensor Decomposition for Latent Variable Models
- 3 Learning Negative Mixtures
 - Negative Mixture of Spherical Gaussians
 - Generalized Tensor Power Method
- 4 Experiments
- 5 Conclusion

Learning Negative Mixtures of Spherical Gaussians

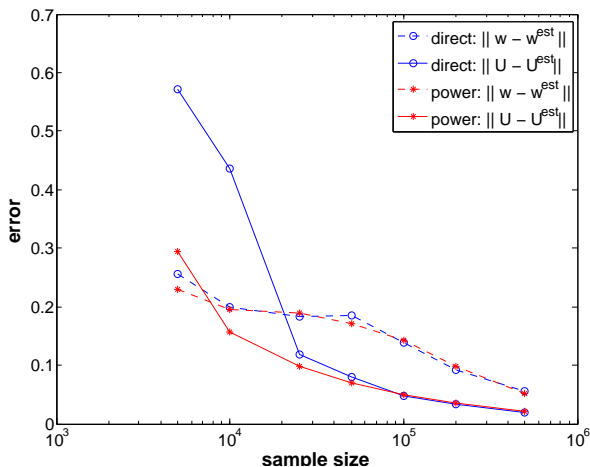


Figure : Negative mixture of 2 spherical Gaussians in dimension 6. Estimation error as a function of the dataset size.

Overview

- 1 Negative Mixtures
 - Definition
 - Negative Mixture of Spherical Gaussians
 - Rational Distribution on Strings
- 2 Learning by Tensor Decomposition
 - Tensors: Preliminaries
 - Learning from Data
 - Tensor Decomposition for Latent Variable Models
- 3 Learning Negative Mixtures
 - Negative Mixture of Spherical Gaussians
 - Generalized Tensor Power Method
- 4 Experiments
- 5 Conclusion



Conclusion

- Contributions

- ▶ Negative mixture models appear naturally
- ▶ Generalized tensor power method and direct recovery via matrix diagonalization
- ▶ Application to negative mixture of spherical Gaussians

- Perspectives

- ▶ Investigate other fields of ML where these models occur
- ▶ Understand the relations between the two methods (and merge them?)
- ▶ Robustness analysis

-  Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2012).
Tensor decompositions for learning latent variable models.
CoRR, abs/1210.7559.
-  Bailly, R. and Denis, F. (2011).
Absolute convergence of rational series is semi-decidable.
Inf. Comput., 209(3):280–295.
-  Hsu, D. and Kakade, S. M. (2013).
Learning mixtures of spherical gaussians: Moment methods and spectral decompositions.
In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 11–20, New York, NY, USA. ACM.