Learning Gaussian mixture models via tensor decomposition

at MLSS 2020

Haolin Chen, in joint work with Luis Rademacher

September 26, 2020

UC Davis

Tensors

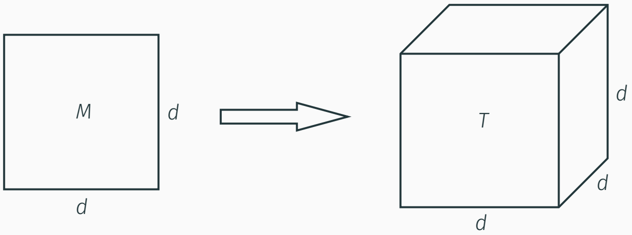
Multi-way arrays:

$$T = \sum_{j_1, j_2, j_3 \in [d]} T_{j_1 j_2 j_3} e_{j_1} \otimes e_{j_2} \otimes e_{j_3}$$

Or multi-linear forms:

$$T : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$$

$$T(x, y, z) = \sum_{j_1, j_2, j_3 \in [d]} T_{j_1 j_2 j_3} x_{j_1} y_{j_2} z_{j_3}.$$



Tensor decomposition

Tensor rank: smallest k such that a tensor can be written in:

$$T = \sum_{i \in [k]} a_i \otimes b_i \otimes c_i.$$

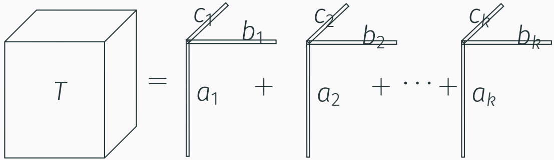
Tensor decomposition

Tensor rank: smallest k such that a tensor can be written in:

$$T = \sum_{i \in [k]} a_i \otimes b_i \otimes c_i.$$

Tensor decomposition: given a rank-k 3-tensor T, find $\{a_i, b_i, c_i, i \in [k]\}$ such that

$$T = \sum_{i \in [k]} a_i \otimes b_i \otimes c_i.$$



Identifiability and algorithm

Theorem ([Kruskal, 1977])

Suppose $T = \sum_{i \in [k]} a_i^{\otimes 3}$ is a symmetric 3-tensor and any d vectors among a_i's are linearly independent, then the decomposition of T is unique if $k \leq 3d/2 - 1$.

Identifiability and algorithm

Theorem ([Kruskal, 1977])

Suppose $T = \sum_{i\in[k]} a_i^{\otimes 3}$ is a symmetric 3-tensor and any d vectors among a_i's are linearly independent, then the decomposition of T is unique if $k \leq 3d/2 - 1$.

Jennrich's algorithm: given $T = \sum_{i\in[k]} a_i \otimes a_i \otimes a_i$, $a_i \in \mathbb{R}^d$.

- goal: recover a_i
- flatten T using random vectors x and y:

$$T_x = T(x,\cdot,\cdot) = \sum_{i\in[k]} (x^\top a_i) a_i a_i^\top \quad T_y = T(y,\cdot,\cdot) = \sum_{i\in[k]} (y^\top a_i) a_i a_i^\top$$

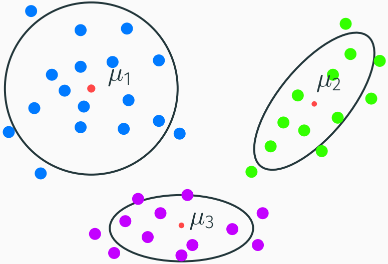
- eigenvectors of $T_x T_y^\dagger$ recover a_i(in the direction), the norm is recoverable by orthogonalizing the tensor.
- works only when a_i's are linearly independent(thus $k \leq d$)

Gaussian mixture model

X is a k-component Gaussian Mixture Model(GMM) in $\mathbb{R}^d$ if

$$X \sim \sum_{i \in [k]} w_i \mathcal{N}(\mu_i, \Sigma^{(i)}),$$

where w_i is the mixing weight s.t. $\sum_{i \in [k]} w_i = 1$, $w_i \in (0, 1)$.

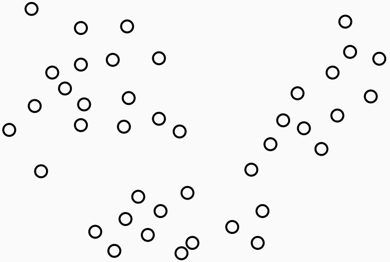


Gaussian mixture model

X is a k-component Gaussian Mixture Model(GMM) in $\mathbb{R}^d$ if

$$X \sim \sum_{i \in [k]} w_i \mathcal{N}(\mu_i, \Sigma^{(i)}),$$

where w_i is the mixing weight s.t. $\sum_{i \in [k]} w_i = 1$, $w_i \in (0, 1)$.



Learning GMMs: estimate the parameters $\{w_i, \mu_i, \Sigma^{(i)}\}$ given finite unlabeled samples.

Learning GMMs via tensor decomposition

Motivation & recipe: method of moments

1. find a tensor encoding the parameters
2. decompose the tensor to recover the parameters

Example: third moment of the discrete distribution $\{w_i, \mu_i : i \in [k]\}$

$$M_3 = \sum_{i \in [k]} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

Learning GMMs via tensor decomposition

Motivation & recipe: method of moments

1. find a tensor encoding the parameters
2. decompose the tensor to recover the parameters

Example: third moment of the discrete distribution $\{w_i, \mu_i : i \in [k]\}$

$$M_3 = \sum_{i \in [k]} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

[Hsu and Kakade, 2013]	Spherical, linearly independent μ_i's
[Anderson et al., 2014]	$O(d^m)$ components with identical and known covariances
[Ge et al., 2015]	$O(\sqrt{d})$ components under smoothed analysis setting
[Hopkins and Li, 2018]	k^γ pairwise separation on μ_i's

Learning GMMs via tensor decomposition

Goal:

1. learn at most $d + c$ Gaussians with identical but unknown covariance matrices and $c \ll d$.

$$X \sim \sum_{i \in [d+c]} w_i \mathcal{N}(\mu_i, \Sigma)$$

2. time, sample complexity: $poly(d)$

Learning GMMs via tensor decomposition

Key idea: third central moment encodes the information we need:

$$T = \sum_{i \in [d+c]} w_i (\mu_i - \bar{\mu})^{\otimes 3}.$$

No existing algorithm can decompose T as we are in the "overcomplete" domain

Learning GMMs via tensor decomposition

Key idea: third central moment encodes the information we need:

$$T = \sum_{i \in [d+c]} w_i (\mu_i - \bar{\mu})^{\otimes 3}.$$

No existing algorithm can decompose T as we are in the "overcomplete" domain

Workaround: 2-steps strategy

1. decompose a subtensor of T
2. deflate T with the reconstructed subtensor
3. decompose the remaining tensor

Flatten T with 2 vectors $x, y \perp \mu_i - \bar{\mu}$ for $i > d$ so that $T(x, \cdot, \cdot), T(y, \cdot, \cdot)$ come from the first d rank one terms in T.

In reality: randomized algorithm proven to stop in polynomial time.

Algorithm outline

Tensor decomposition algorithm: $T = \sum_{i \in [d+c]} a_i \otimes a_i \otimes a_i$.

Input: 3-tensor T, error tolerance ϵ

repeat:

1. pick x, y uniformly at random on the unit sphere
2. invoke Jennrich's algorithm with x, y
3. deflate recovered components from T
4. pick x', y' uniformly at random on the unit sphere
5. invoke Jennrich's algorithm with x', y' on the remaining tensor

until: reconstruction error $\leq \epsilon$

Output: $\tilde{a}_i$ such that $\| \sum_{i \in [d+c]} \tilde{a}_i^{\otimes 3} - a_i^{\otimes 3} \|_F \leq \epsilon$

Algorithm outline

Gaussian mixture learning: recall 3rd central moment

$$T = \sum_{i \in [d+c]} w_i (\mu_i - \bar{\mu})^{\otimes 3}$$

Input: 1st and 2nd moments, 3rd central moment T, error tolerance ϵ

1. decompose T with tolerance ϵ
2. decouple mixing weights w_i and $\mu_i - \bar{\mu}$ from $w_i \|\mu_i - \bar{\mu}\|$
3. recover μ_i, Σ using other moments

Output: estimated parameters $\tilde{w}_i, \tilde{\mu}_i, \tilde{\Sigma}$

Proof idea at a glance

Provable results:

1. *poly*(d) sample complexity
2. robust to $1/\mathit{poly}(d)$ error
3. full algorithm expected to end in *poly*(d) time.

Proof ideas:

1. finite higher order moments guarantees the polynomial sample complexity;
2. use standard matrix perturbation theory on eigendecomposition to show the robustness on error
3. high dimensional probability bounds guarantee with positive probability we could find some “magic” vectors satisfying our requirements.

Summary and more

Summary:

1. an overcomplete tensor decomposition algorithm
2. a Gaussian mixture learning algorithm that generalizes to more general mixture models

Future directions:

1. tensor decomposition with $O(d^n)$ components
2. mixture learning of general Gaussians

Anderson, J., Belkin, M., Goyal, N., Rademacher, L., and Voss, J. (2014).
The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures.
In *Conference on Learning Theory*, pages 1135–1164.

Ge, R., Huang, Q., and Kakade, S. M. (2015).
Learning mixtures of gaussians in high dimensions.
In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 761–770. ACM.

Hopkins, S. B. and Li, J. (2018).
Mixture models, robustness, and sum of squares proofs.
In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM.

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*, pages 11–20. ACM.

Kruskal, J. B. (1977).
Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics.
Linear algebra and its applications, 18(2):95–138.