Illustration by Chris Brigman

Tensor Moments of Gaussian Mixture Models: Theory and Applications

Tammy Kolda
MathSci.ai

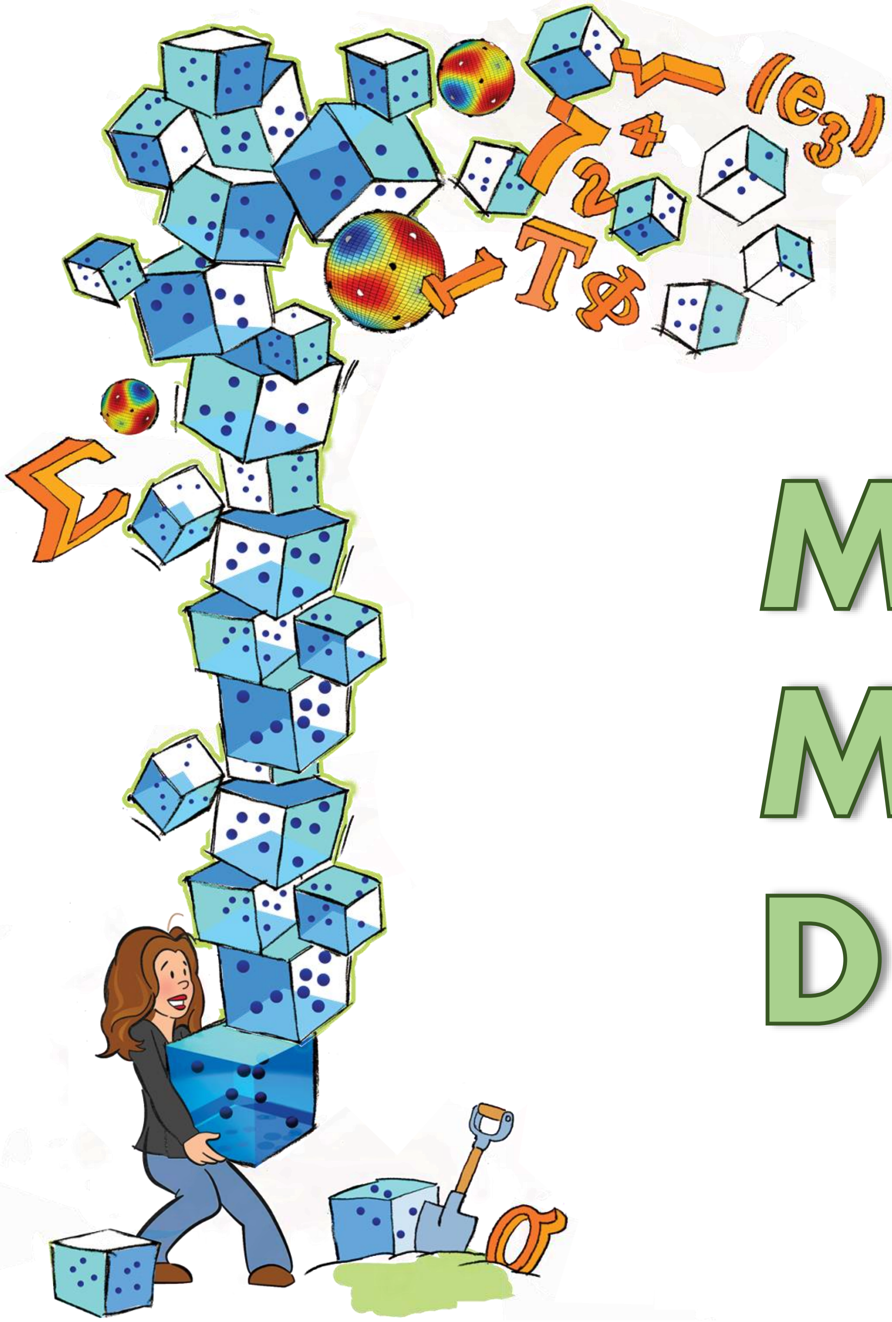
Joe Kileel and João M. Pereira
University of Texas, Austin

Amazing Coauthors

Joe Kileel
Asst Professor
Math Dept
U Texas, Austin

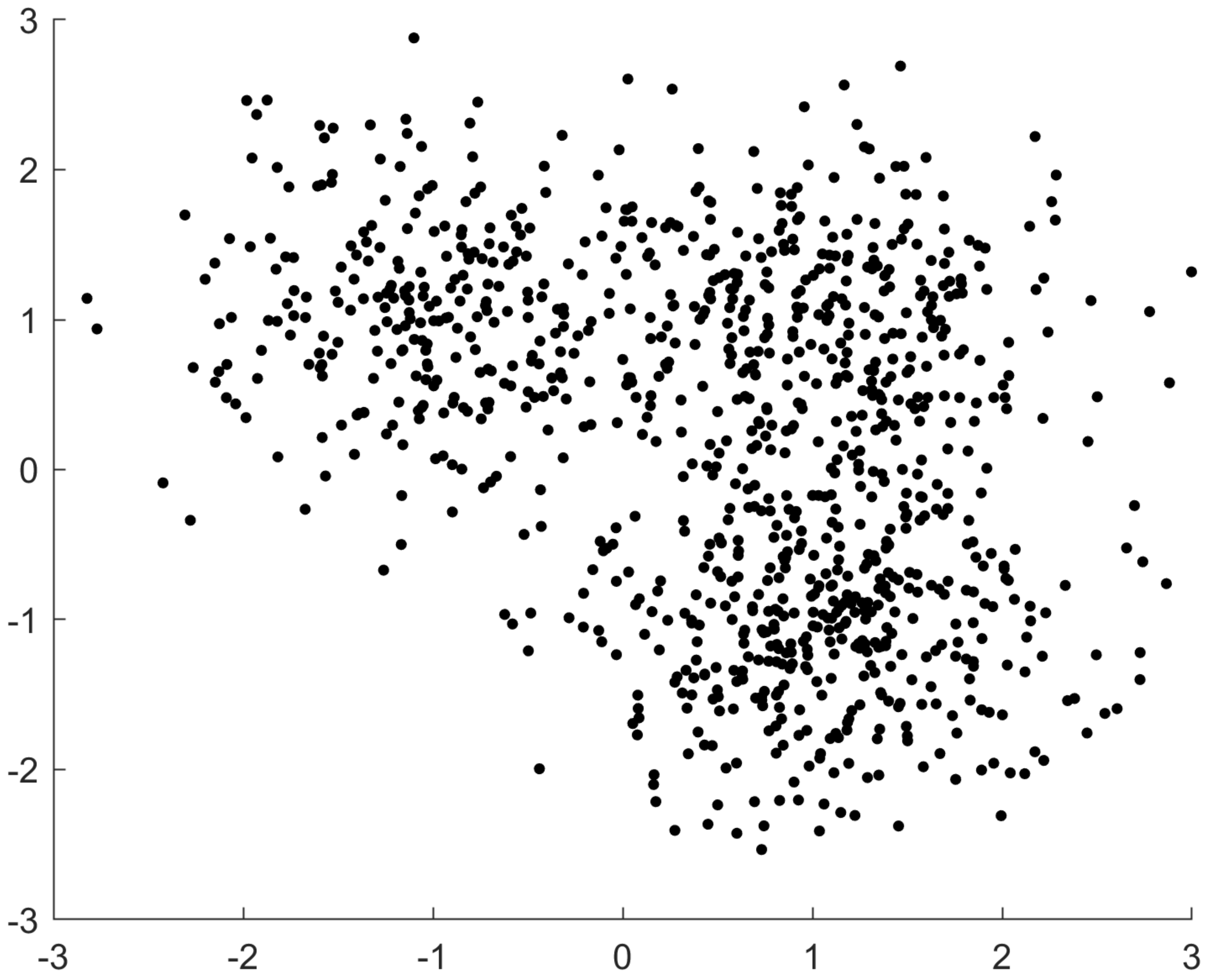
João M. Pereira
Postdoc
Math Dept
U Texas, Austin

(starting summer 2022)
Asst Professor
Instituto Nacional de Matemática Pura e Aplicada (IMPA)
Brazil



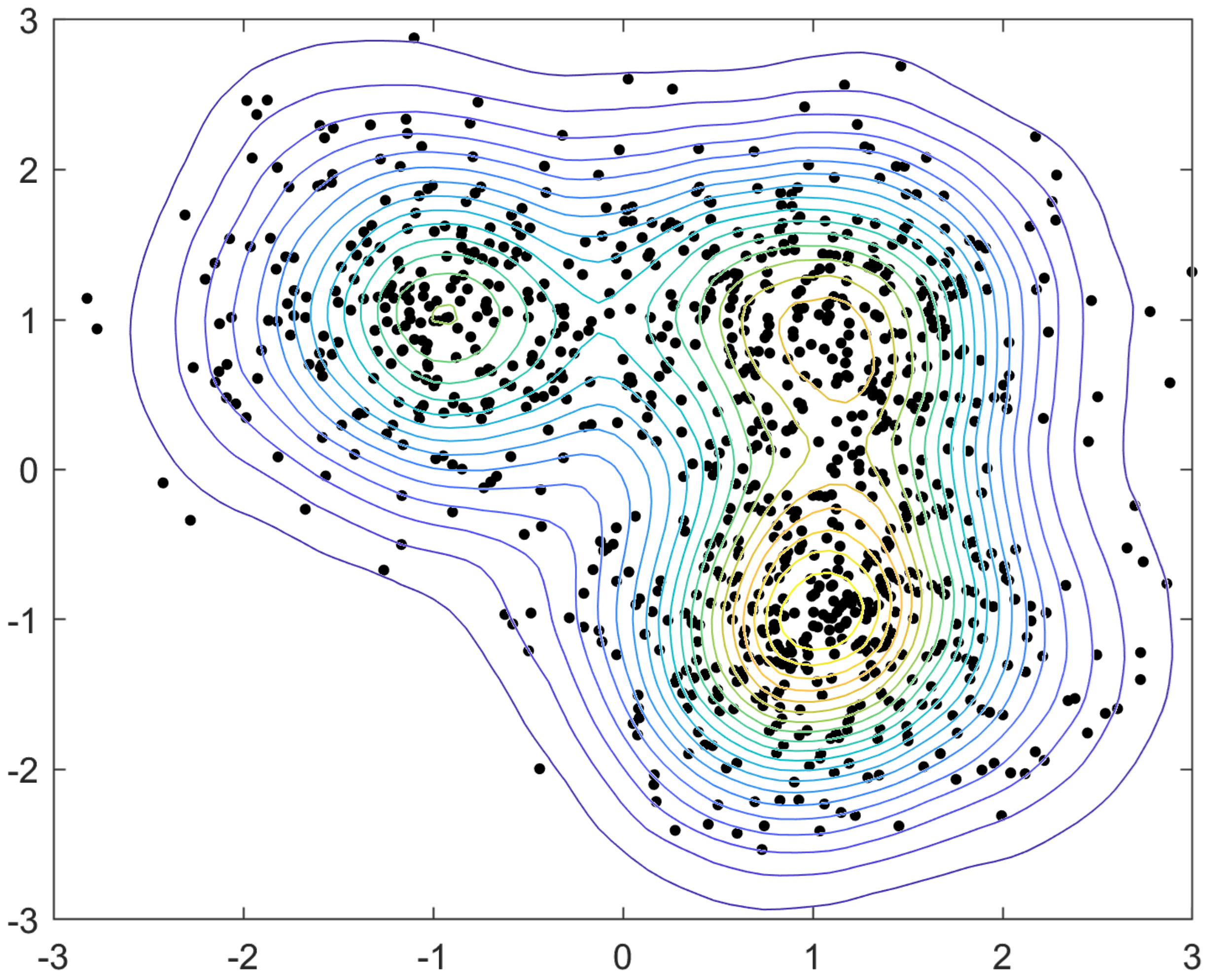
Motivation: Making Sense of Data via Models

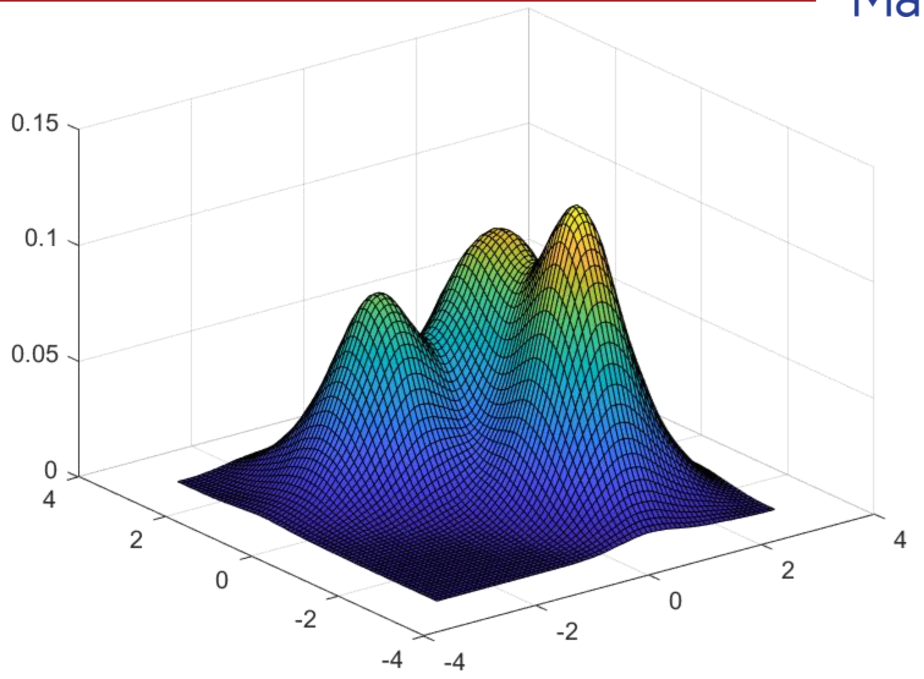
Motivation: Making Sense of Data

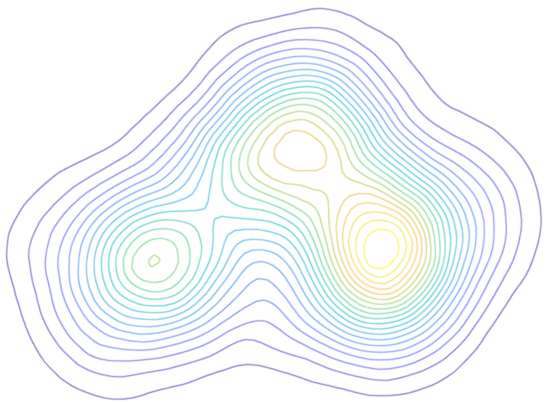


Illustrative Example:
1000 datapoints
(x_1, x_2)

Technique: Create Model of Probability Density from Datapoints







Gaussian Distribution ("Bell curve")

MathSci.ai

Probability Distribution Function (pdf)

$$\frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}}$$

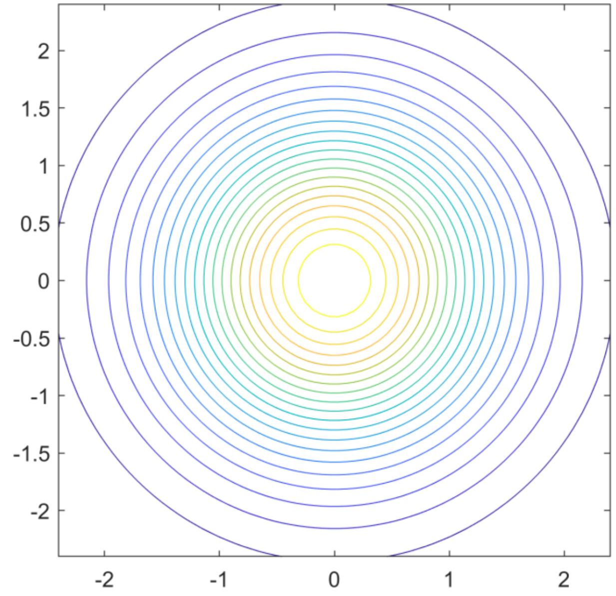
$$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

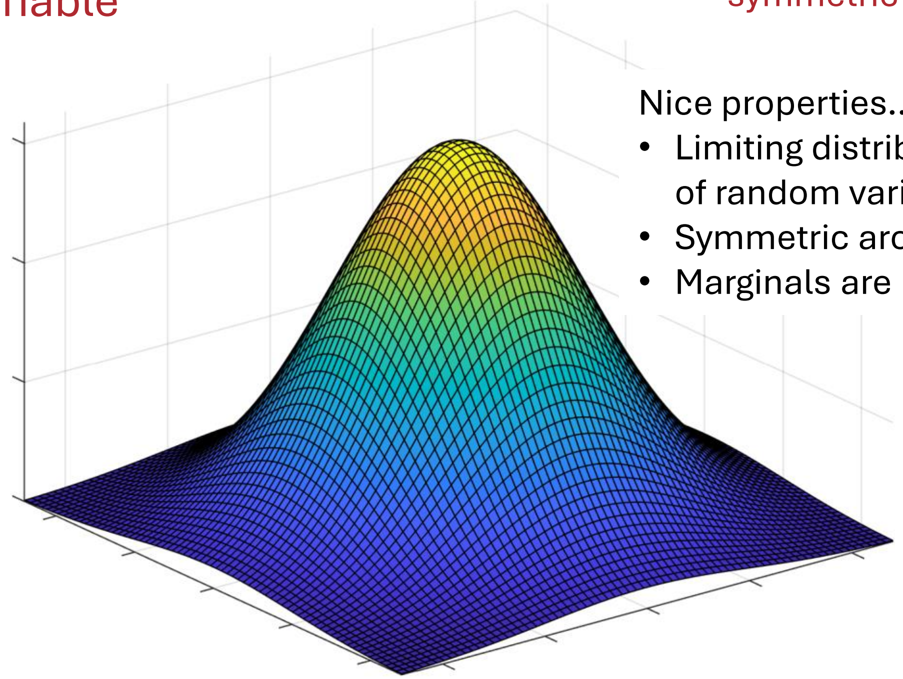
n-dimensional random variable

mean or center $\boldsymbol{\mu} \in \mathbb{R}^n$

covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$
symmetric positive definite

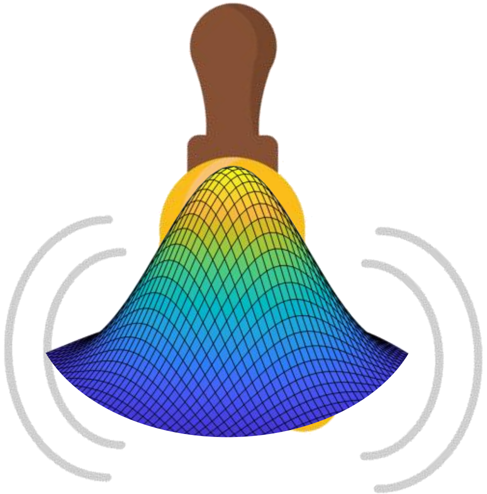
aka "Normal" distribution





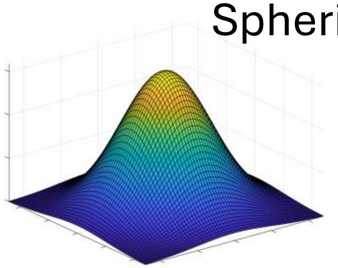
Nice properties...
- Limiting distribution of a sum of random variables
- Symmetric around the mean
- Marginals are normal

Bell Curve

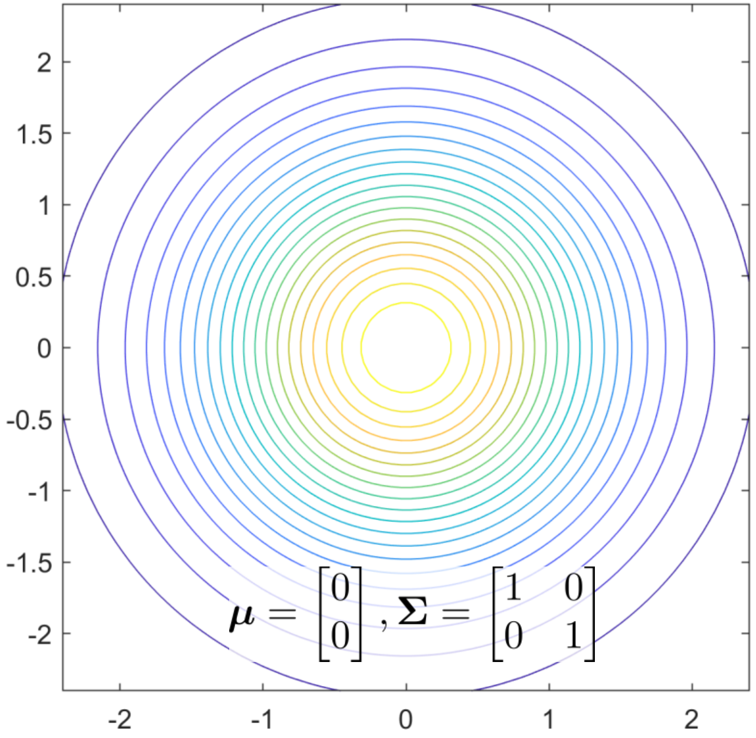


Covariance for Gaussian Distribution

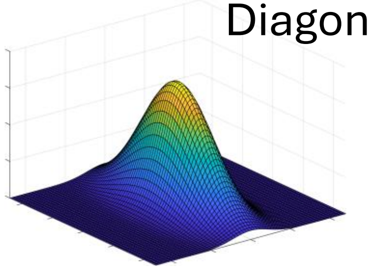
Spherical or Isotropic



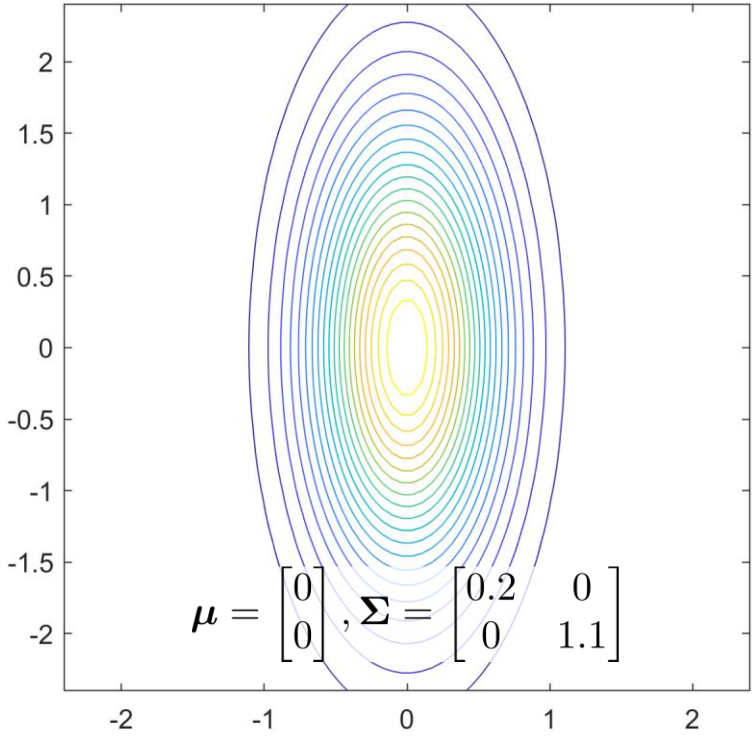
$$\mathbf{\Sigma} = \sigma^2 \mathbf{I}$$



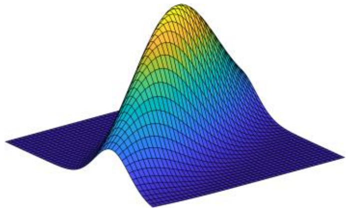
Diagonal or Axis-Aligned

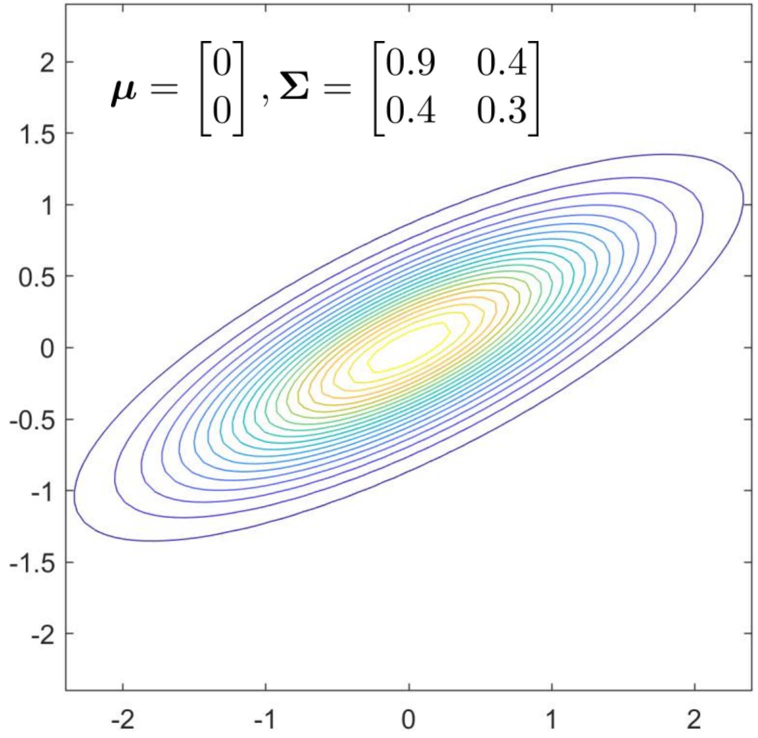


$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

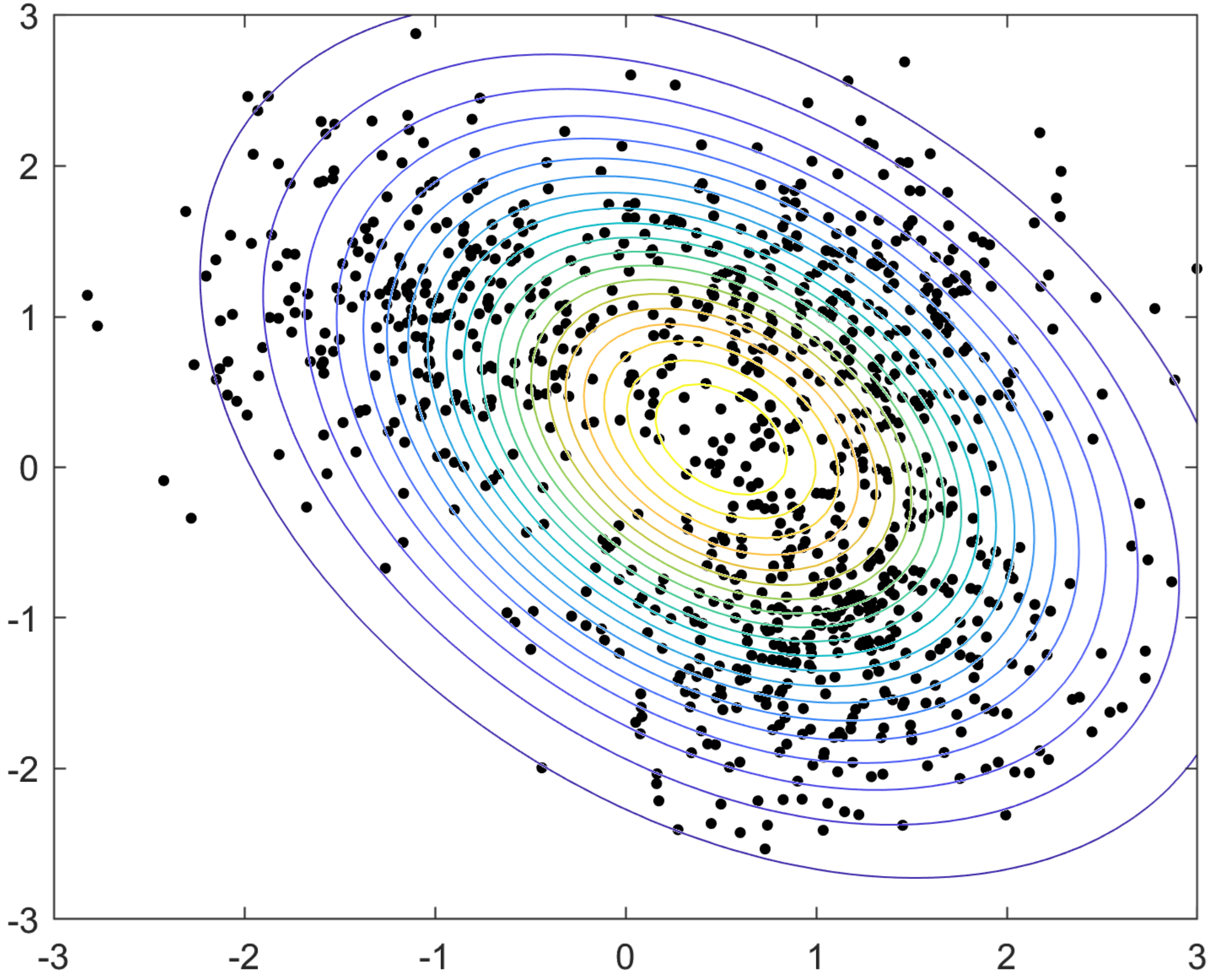


General

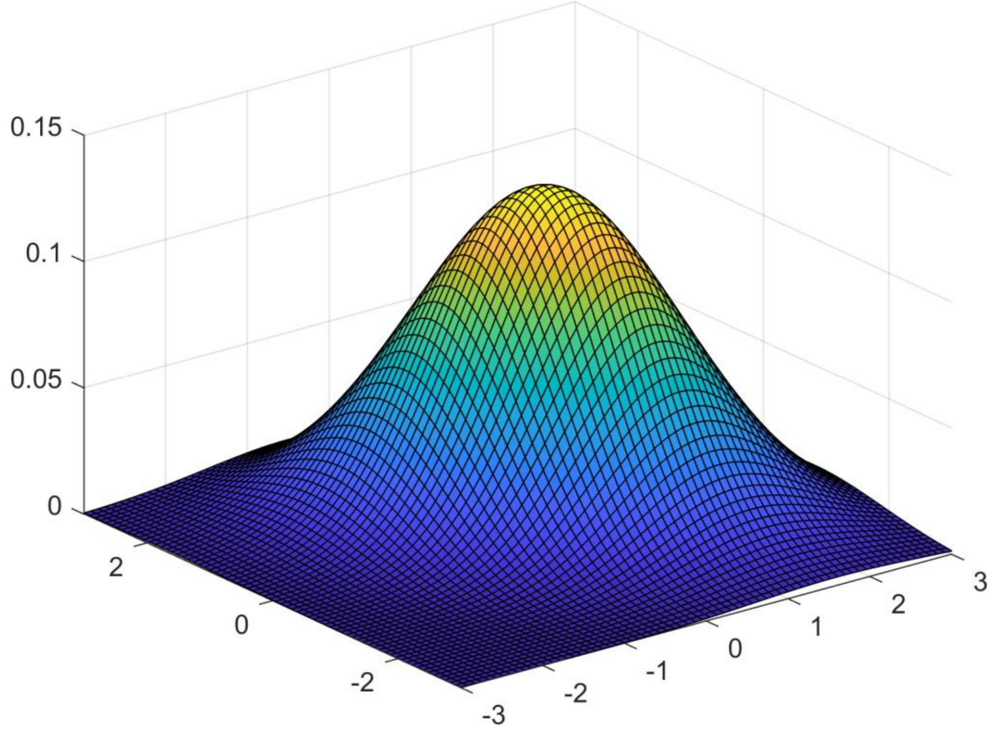




Gaussian Model of Data



Parameters: $\boldsymbol{\mu}, \boldsymbol{\Sigma}$



Gaussian Mixture Model (GMM)

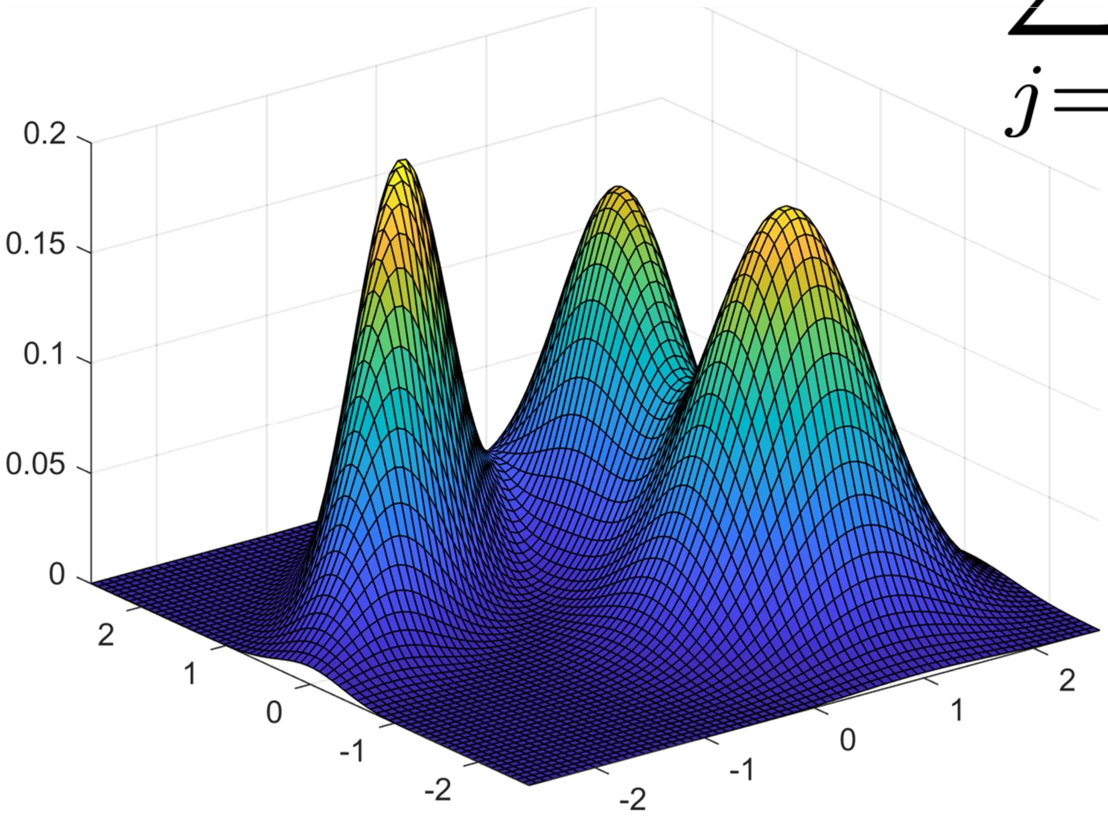
MathSci.ai

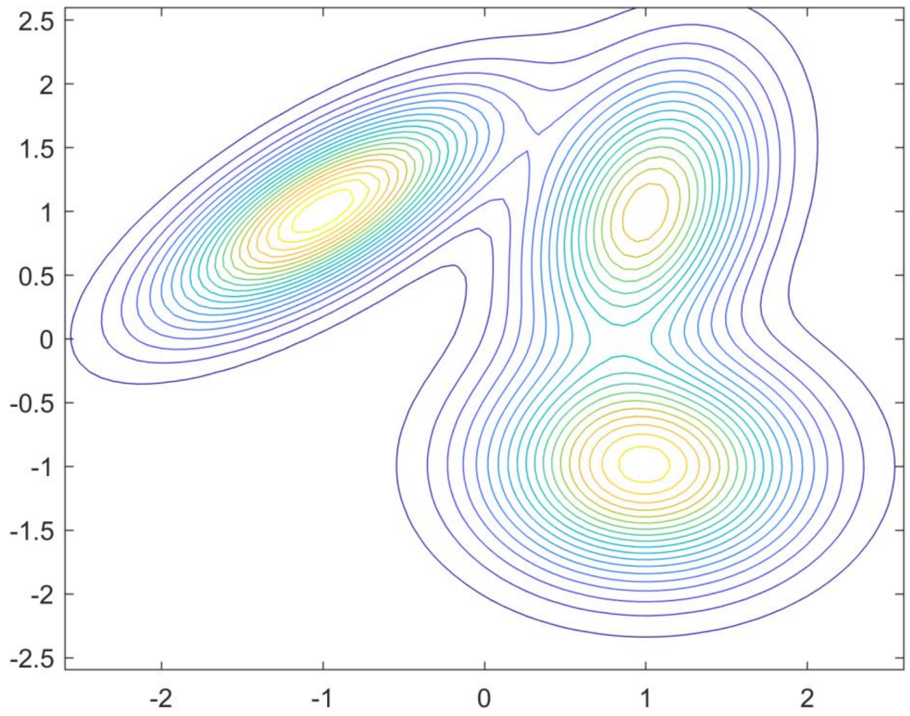
$$X \sim \overbrace{\sum_{j=1}^{m} \lambda_j \, \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Number of Components → m

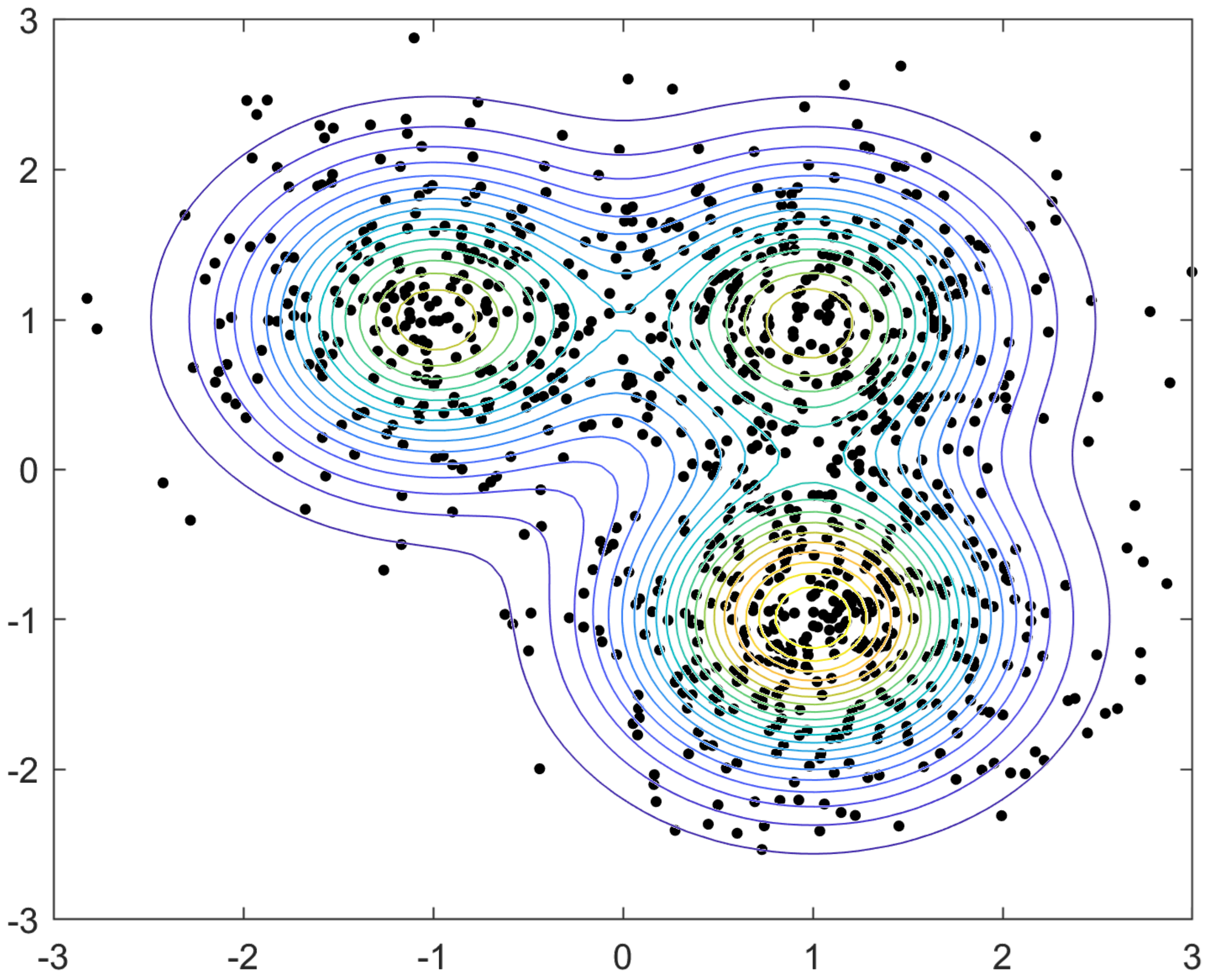
Component j

Proportion in
jth component

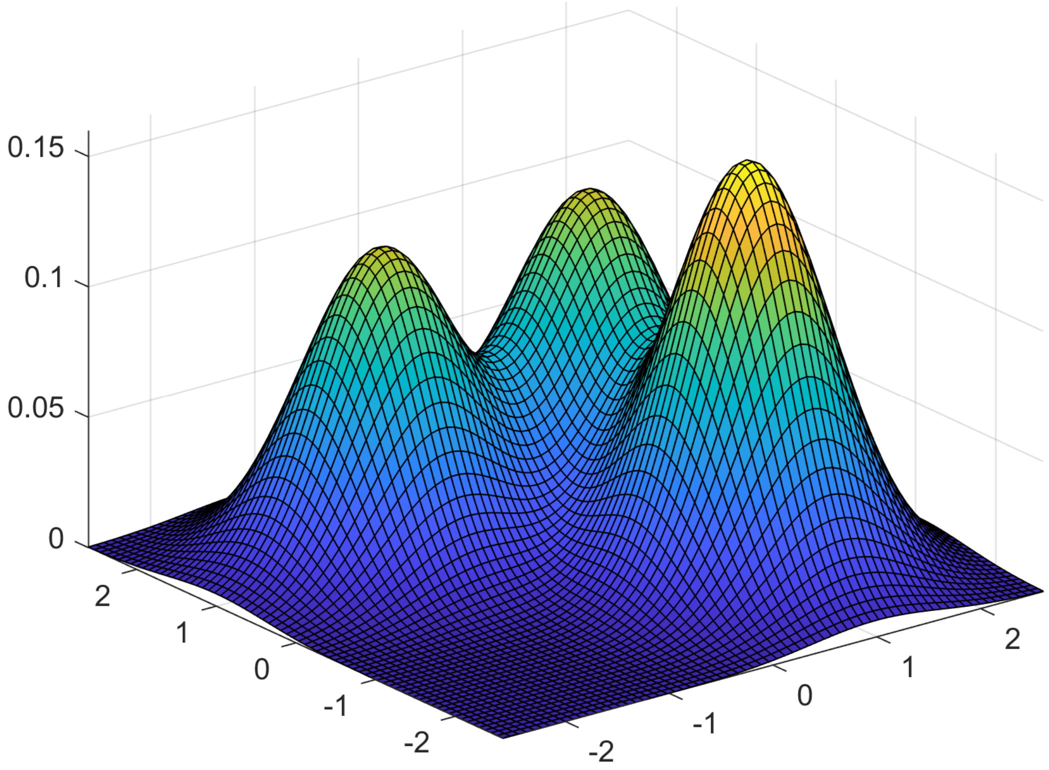




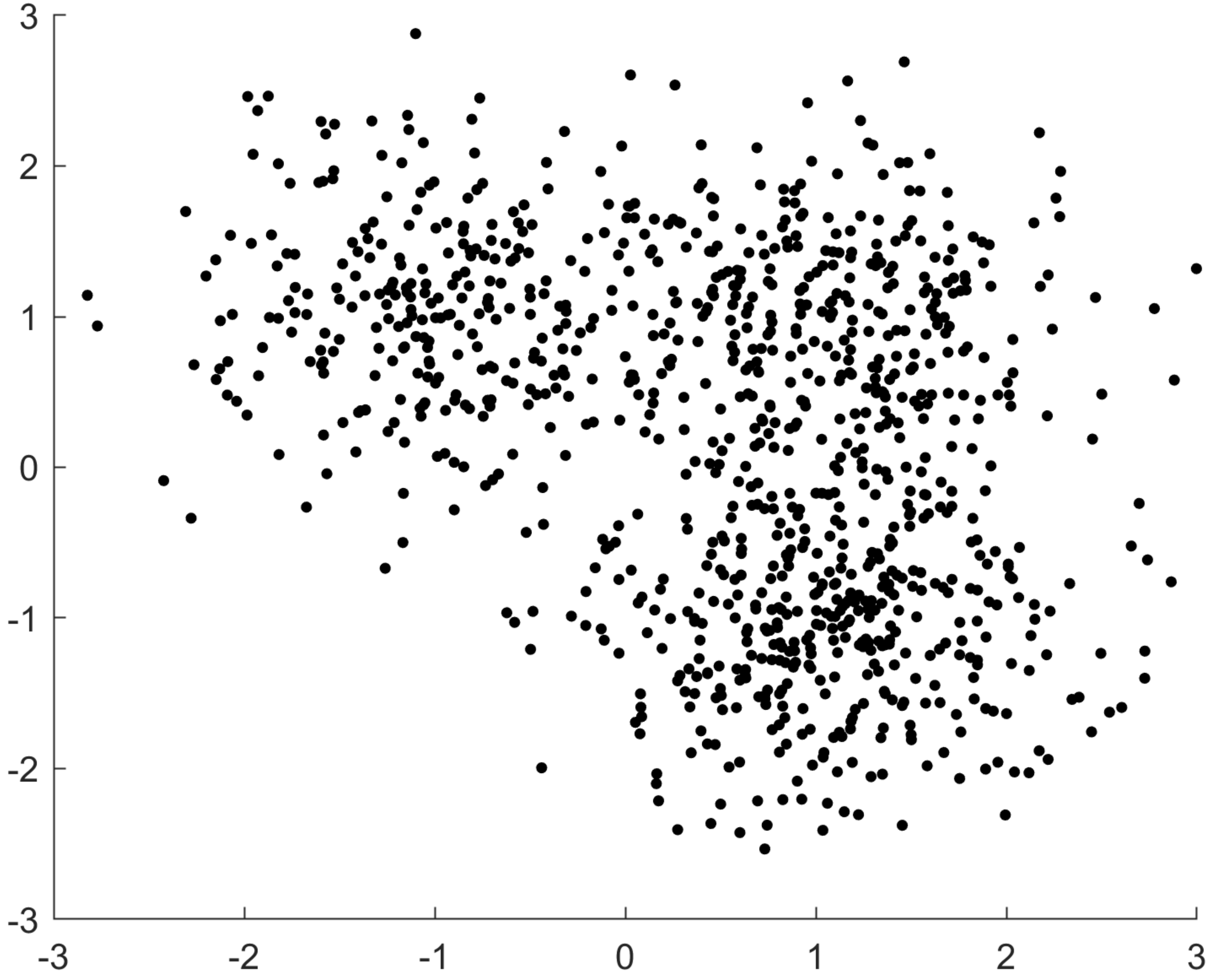
GMM with Three Components



Parameters: $\{\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1:m}$



Motivation: Making Sense of Data



Illustrative Example:
1000 datapoints
(x_1, x_2)

Real-World Scenario:
10,000+ datapoints
$(x_1, x_2, \dots, x_{500})$

Many more datapoints

Much higher dimension

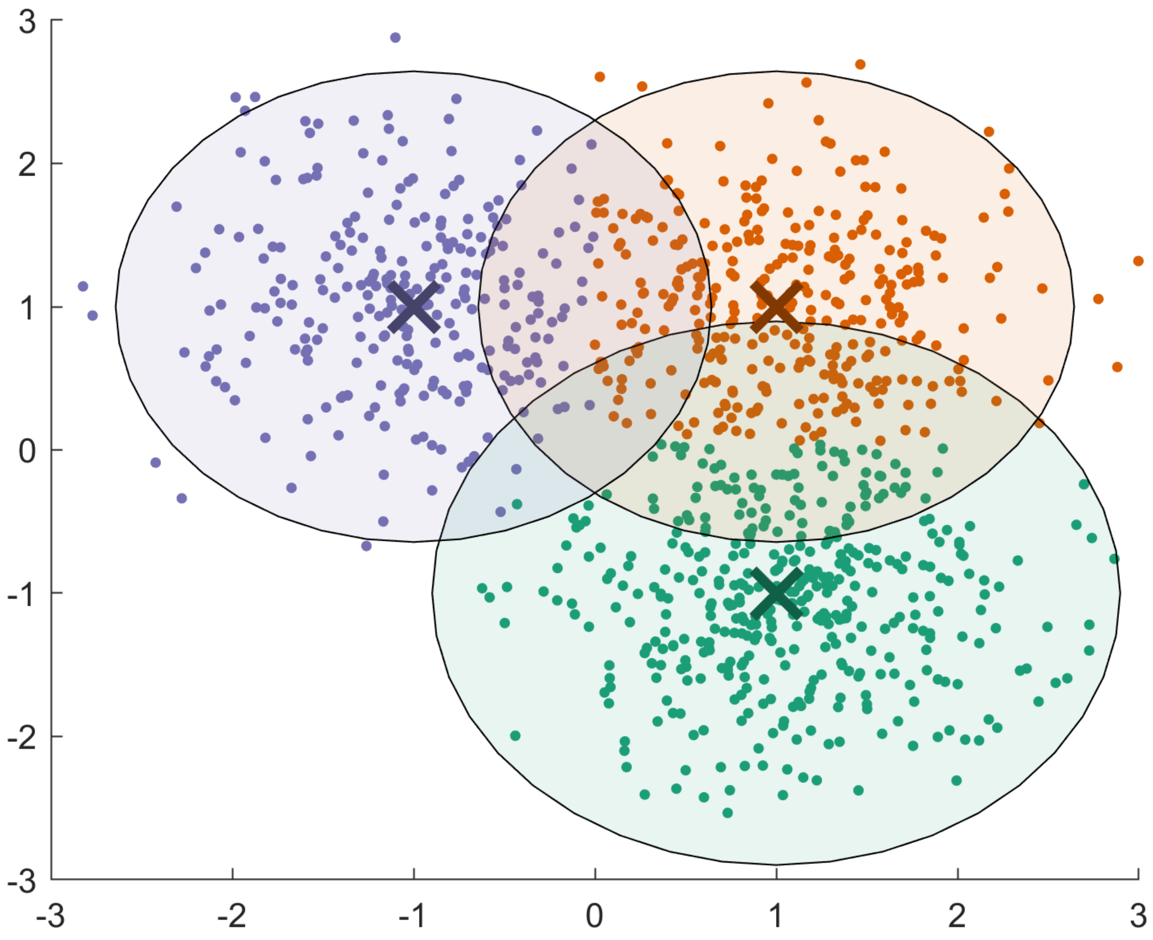
Applications: Gaussian Mixtures

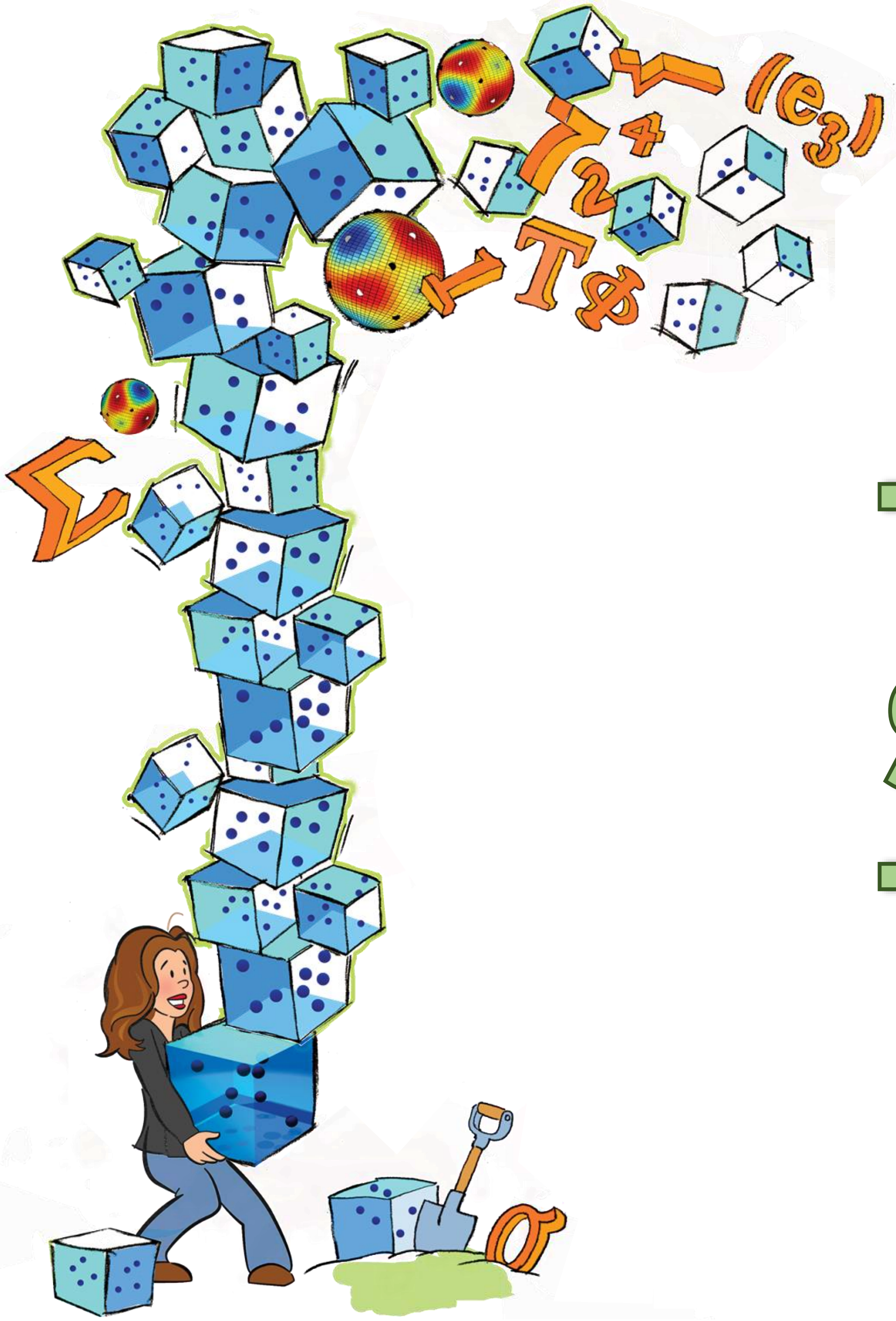
MathSci.ai

Density Estimation

Clustering

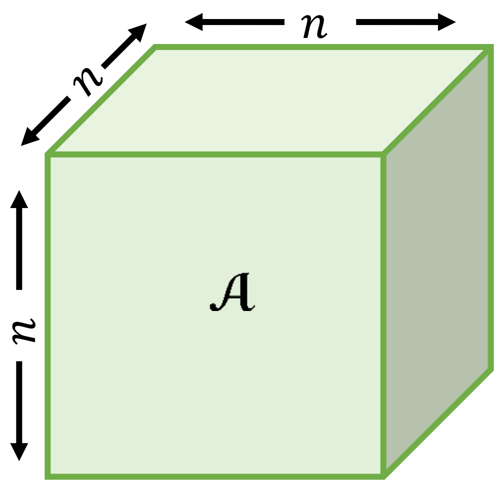
Anomaly Detection





Tensors and Symmetric Tensors

Tensors are Multi-dimensional Arrays



3-way tensor

d = **order** of the tensor (the number of ways or modes)

For this talk, all modes have the same size: n

n^d = number of entries for d-way tensor of dimension n

Curse of Dimensionality

a_{ijk} = (i, j, k) entry of 3-way tensor $\mathcal{A}$

A tensor is **symmetric** if its entries are invariant under permutation, i.e.,

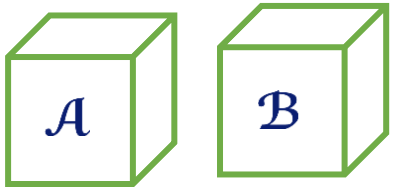
$$a_{ijk} = a_{ikj} = a_{jik} = a_{jik} = a_{jki} = a_{kij} = a_{kji}$$

Curse of notation…

$(i_1, i_2, \ldots, i_d)$ = **index** into tensor, $i_k \in \{1, \ldots, n\}$ for $k = 1, 2, \ldots, d$

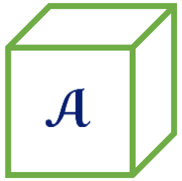
Tensor Norm & Inner Product

Inner Product

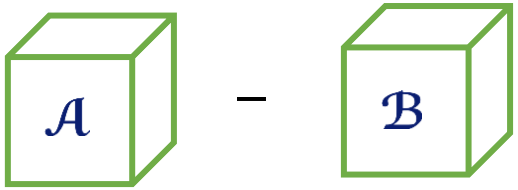


$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk} b_{ijk}$$

Norm



$$\|\mathcal{A}\|^2 = \langle \mathcal{A}, \mathcal{A} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk}^2$$



$$\|\mathcal{A} - \mathcal{B}\|^2 = \|\mathcal{A}\|^2 + \|\mathcal{B}\|^2 - 2\langle \mathcal{A}, \mathcal{B} \rangle$$

Symmetric Tensor Outer Product

MathSci.ai

$\mathbf{b} \in \mathbb{R}^n$

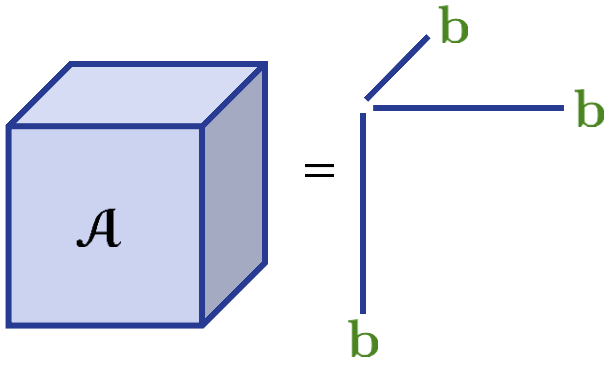
$$\mathbf{A} = \mathbf{b}^{\otimes 2} = \mathbf{b} \otimes \mathbf{b} \in \mathbb{R}^{n \times n} \quad \Longleftrightarrow \quad a_{ij} = b_i b_j$$

$$\mathcal{A} = \mathbf{b}^{\otimes 3} = \mathbf{b} \otimes \mathbf{b} \otimes \mathbf{b} \in \mathbb{R}^{n \times n \times n} \quad \Longleftrightarrow \quad a_{ijk} = b_i b_j b_k$$

$$\mathcal{A} = \mathbf{b}^{\otimes 4} = \mathbf{b} \otimes \mathbf{b} \otimes \mathbf{b} \otimes \mathbf{b} \in \mathbb{R}^{n \times n \times n \times n} \quad \Longleftrightarrow \quad a_{ijk\ell} = b_i b_j b_k b_\ell$$

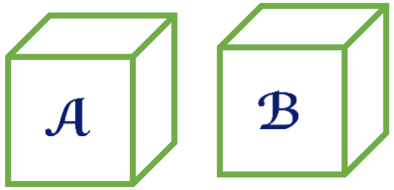
$$\mathcal{A} = \mathbf{b}^{\otimes d} = \underbrace{\mathbf{b} \otimes \cdots \otimes \mathbf{b}}_{d \text{ times}} \in \mathbb{R}^{n \times \cdots \times n} \quad \Longleftrightarrow \quad a_{i_1 \cdots i_d} = b_{i_1} \cdots b_{i_d}$$

Visualization of 3-way Outer Product



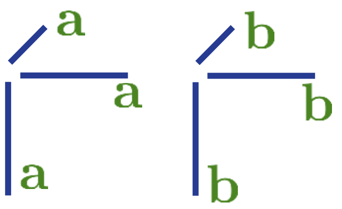
Outer Products, Inner Products, and Norms

Inner Product



$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk} b_{ijk}$$

Inner Product of Outer Products

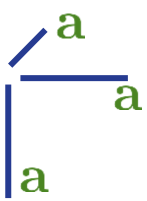


$$\langle \mathbf{a}^{\otimes 3}, \mathbf{b}^{\otimes 3} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} (a_i a_j a_k)(b_i b_j b_k)$$

$$= \left(\sum_{i=1}^{n} a_i b_i \right) \left(\sum_{j=1}^{n} a_j b_j \right) \left(\sum_{k=1}^{n} a_k b_k \right) = \langle \mathbf{a}, \mathbf{b} \rangle^3$$

$$\langle \mathbf{a}^{\otimes d}, \mathbf{b}^{\otimes d} \rangle = \langle \mathbf{a}, \mathbf{b} \rangle^d$$

Norm



$$\| \mathbf{a}^{\otimes d} \|^2 = \langle \mathbf{a}^{\otimes d}, \mathbf{a}^{\otimes d} \rangle = \langle \mathbf{a}, \mathbf{a} \rangle^d = \| \mathbf{a} \|^{2d}$$

Approximate Method of Moments for Gaussian Mixtures

See Sherman & K (2020)

Moments of a Multivariate Random Variable

Higher moments capture interactions between the variables

Random variable: $X \in \mathbb{R}^n$

First moment: $\mathcal{M}^{(1)} = \mathbb{E}(X)$ $\quad$ $\mathcal{M}^{(1)}(i) = \mathbb{E}(X_i)$

Second moment: $\mathcal{M}^{(2)} = \mathbb{E}(X^{\otimes 2})$ $\quad$ $\mathcal{M}^{(2)}(i,j) = \mathbb{E}(X_i X_j)$

Third moment: $\mathcal{M}^{(3)} = \mathbb{E}(X^{\otimes 3})$ $\quad$ $\mathcal{M}^{(3)}(i,j,k) = \mathbb{E}(X_i X_j X_k)$

dth moment: $\mathcal{M}^{(d)} = \mathbb{E}(X^{\otimes d})$

dth moment is a symmetric tensor or order d

Moments Define a Distribution

Higher moments capture interactions between the variables

Random variable: $X \in \mathbb{R}^n$

First moment: $\mathcal{M}^{(1)} = \mathbb{E}(X)$

Second moment: $\mathcal{M}^{(2)} = \mathbb{E}(X^{\otimes 2})$

Third moment: $\mathcal{M}^{(3)} = \mathbb{E}(X^{\otimes 3})$

dth moment: $\mathcal{M}^{(d)} = \mathbb{E}(X^{\otimes d})$

"Method of Moments" matches empirical and model moments to estimate the parameters of a distribution.

We focus primarily on matching just the dth moment

Gaussian Mixture Model: Small Spherical Covariance

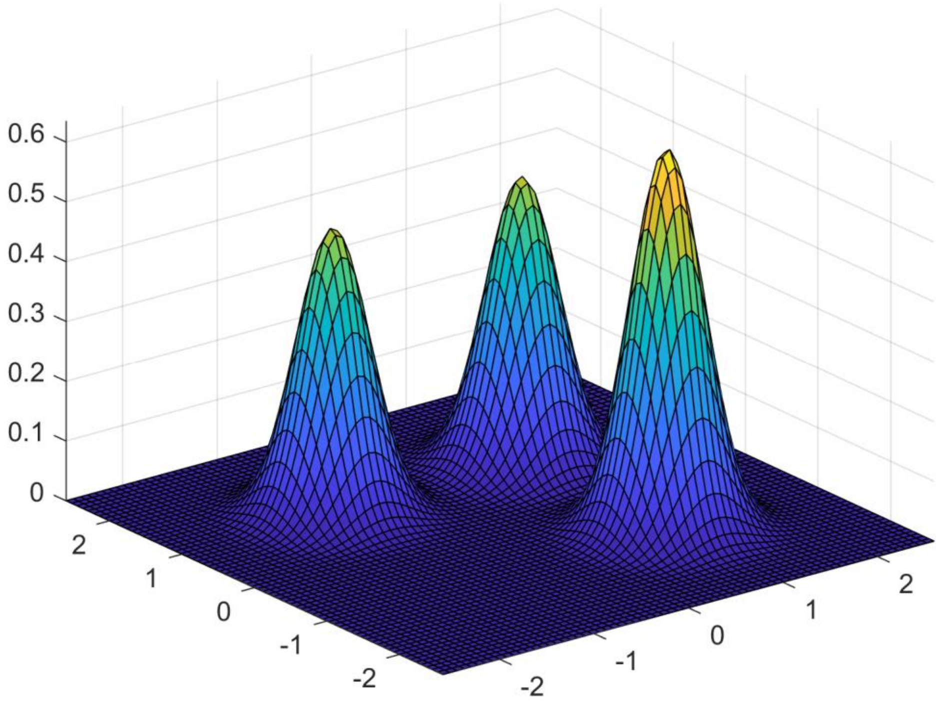
$$X \sim \sum_{j=1}^{m} \lambda_j \, \mathcal{N}(\boldsymbol{\mu}_j, \sigma^2 \mathbf{I})$$

Moment of the Model:

$$\mathfrak{M}^{(d)} = \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d} + \mathcal{O}(\sigma^2)$$

Approximate Moment of the Model:

$$\widetilde{\mathfrak{M}}^{(d)} = \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d}$$



Fitting the dth Moment

MathSci.ai

Given r observations:

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\} \subset \mathbb{R}^n$$

Find m weights and mean vectors:

$$\{\lambda_1, \lambda_2, \ldots, \lambda_m\} \subset \mathbb{R}$$

$$\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_m\} \subset \mathbb{R}^n$$

ignoring covariance $\sigma^2\mathbf{I}$

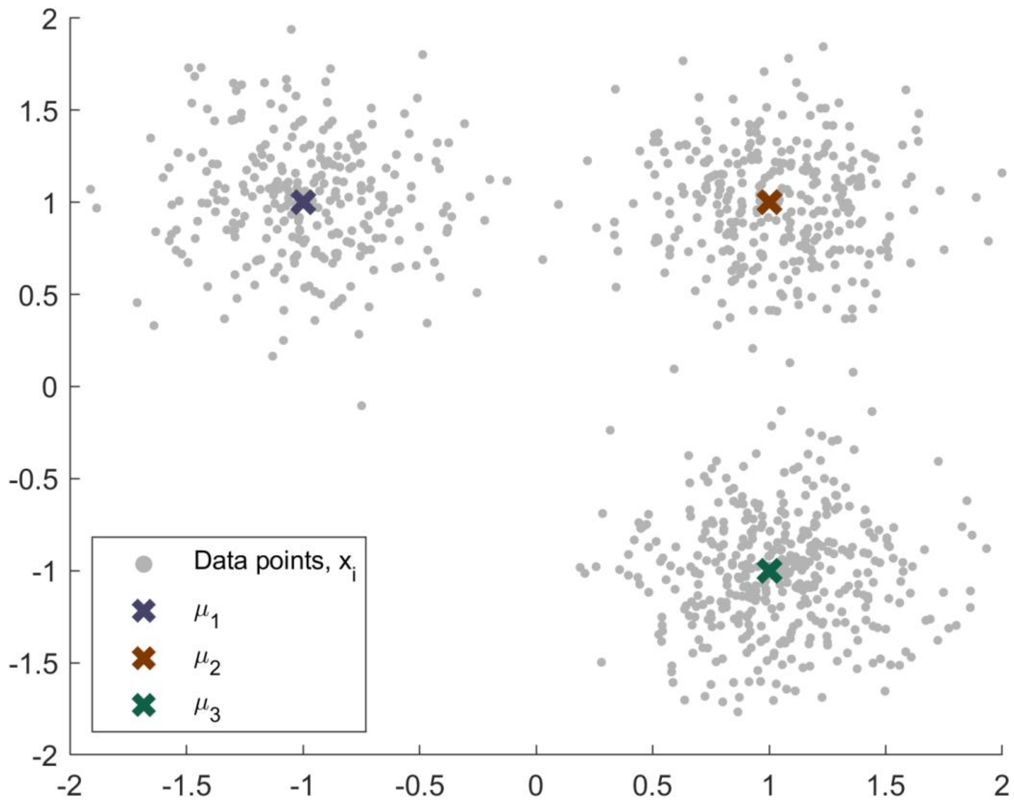
Such that...

$$\sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d} \approx \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d}$$

Approximate Model Moment

Empirical or Sample Moment

Example: $r = 1000, m = 3$



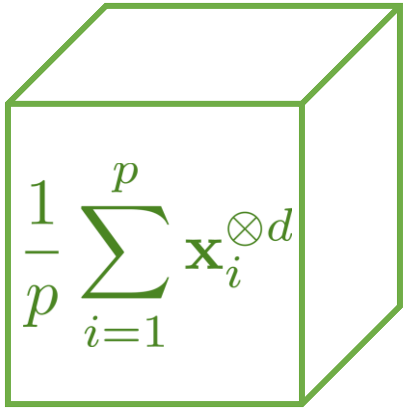
Optimization Formulation: Symmetric Tensor Decomposition

Optimization Parameters

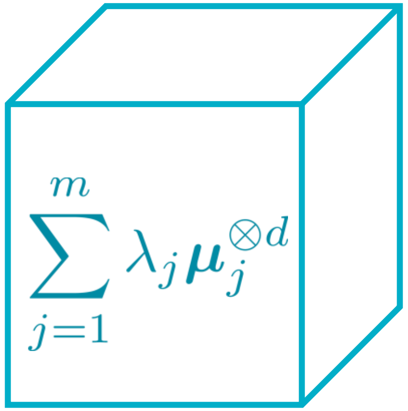
$$\theta = \{\lambda_j, \boldsymbol{\mu}_j\}_{j=1}^{m} \qquad \min_{\theta} f(\theta) \equiv \left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} - \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d} \right\|^2$$

Problem: Forming and storing the empirical moment tensors costs $\mathcal{O}(pn^d)$ operations and $\mathcal{O}(n^d)$ storage

empirical moment tensor



approximate model moment tensor



Problem: Forming and storing the approximate model moment tensor costs $\mathcal{O}(mn^d)$ operations and $\mathcal{O}(n^d)$ storage

But there is only $\mathcal{O}(pn)$ data and $\mathcal{O}(mn)$ parameters, so there is room for efficiency

Optimization Formulation as Inner Products

MathSci.ai

$$\min_\theta f(\theta) \equiv \left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} - \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d} \right\|^2$$

$\theta = \{ \lambda_j, \boldsymbol{\mu}_j \}_{j=1}^{m}$

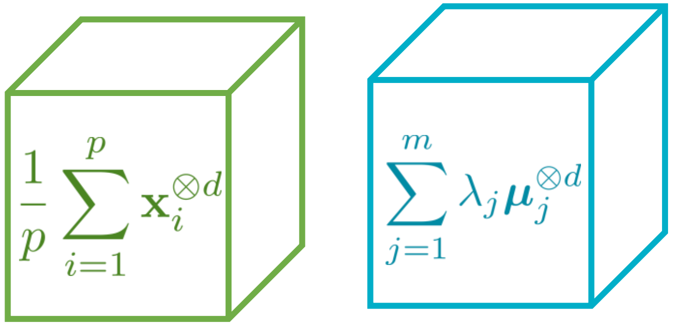
$$f(\theta) = \underbrace{\cancel{\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} \right\|^2}}_{\text{constant}} + \left\| \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d} \right\|^2 - 2 \left\langle \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d}, \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d} \right\rangle$$

$$f(\theta) = C + \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j \left\langle \boldsymbol{\mu}_i^{\otimes d}, \boldsymbol{\mu}_j^{\otimes d} \right\rangle - \frac{2}{p} \sum_{i=1}^{p} \sum_{j=1}^{m} \lambda_j \left\langle \mathbf{x}_i^{\otimes d}, \boldsymbol{\mu}_j^{\otimes d} \right\rangle$$

$$f(\theta) = C + \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle^d - \frac{2}{p} \sum_{i=1}^{p} \sum_{j=1}^{m} \lambda_j \langle \mathbf{x}_i, \boldsymbol{\mu}_j \rangle^d$$

Dot products!

Casting as Optimization Problem



$$f(\theta) = C + \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle^d - \frac{2}{p}\sum_{i=1}^{p}\sum_{j=1}^{m} \lambda_j \langle \mathbf{x}_i, \boldsymbol{\mu}_j \rangle^d$$

$$\theta = \{\, \lambda_j, \boldsymbol{\mu}_j \,\}_{j=1}^{m}$$

- Never need to form empirical or approximate model moments explicitly, overcoming curse of dimensionality
- Function can be calculated with only dot products, total work $\mathcal{O}(m^2 n + pmn)$ and $\mathcal{O}(mn + pn)$ storage, versus $\mathcal{O}\left(mn^d + pn^d\right)$ as originally formulated
- Gradients equally efficient to calculate, via chain rule
- *Issue:* Inherent scaling problem (will come back to this later)
- Easy stochastic function and gradient if number of samples (p) is large…

$$\tilde{f}(\theta) = C + \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle^d - \frac{2}{|\Omega|}\sum_{i\in\Omega}\sum_{j=1}^{m} \lambda_j \langle \mathbf{x}_i, \boldsymbol{\mu}_j \rangle^d$$

But the Approximation is Biased

$$X \sim \sum_{j=1}^{m} \lambda_j \, \mathcal{N}(\boldsymbol{\mu}_j, \sigma^2 \mathbf{I})$$

$$\mathbb{E}(X^{\otimes d}) = \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d} + \mathcal{O}(\sigma^2)$$

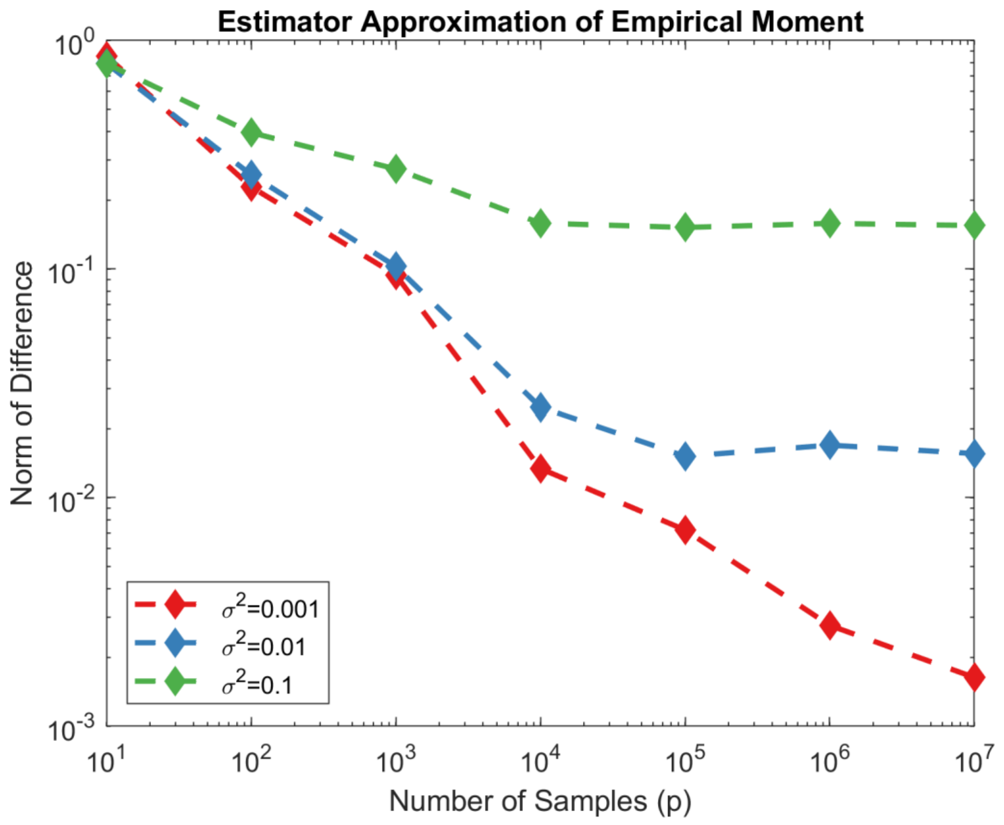
approximate model moment tensor

$$\widetilde{\mathcal{M}}^{(d)} = \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes d}$$

empirical moment tensor

$$\hat{\mathcal{M}}^{(d)} = \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d}$$

$$\left\| \hat{\mathcal{M}}^{(d)} - \widetilde{\mathcal{M}}^{(d)} \right\| = \left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes 3} - \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes 3} \right\|$$



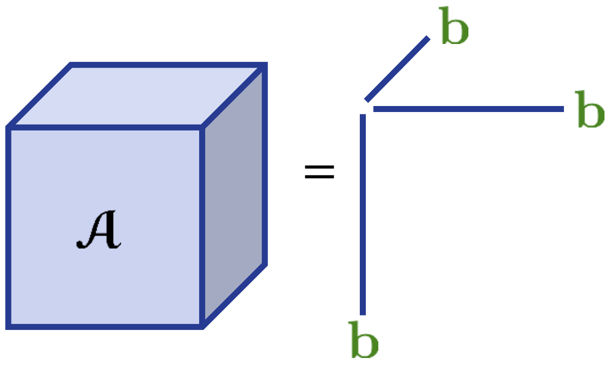
MathSci.ai

More Tensors and Symmetric Tensors

More Tensor Products

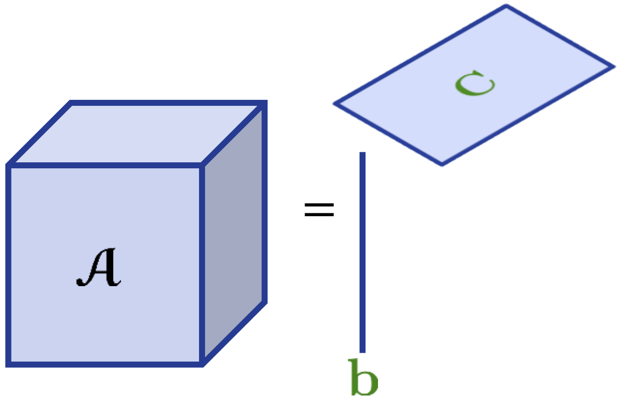
$\mathbf{b} \in \mathbb{R}^n \qquad \mathbf{C} \in \mathbb{R}^{n \times n}$

$\mathcal{A} = \mathbf{b}^{\otimes 3} = \mathbf{b} \otimes \mathbf{b} \otimes \mathbf{b} \in \mathbb{R}^{n \times n \times n} \iff a_{ijk} = b_i b_j b_k$



$\mathcal{A} = \mathbf{b} \otimes \mathbf{C} \in \mathbb{R}^{n \times n \times n} \iff a_{ijk} = b_i c_{jk}$

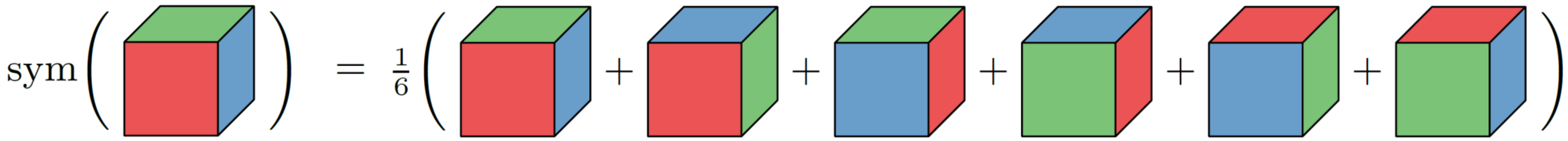
(not symmetric!)



$\mathcal{A} = \mathbf{b}^{\otimes 2} \otimes \mathbf{C} \in \mathbb{R}^{n \times n \times n \times n} \iff a_{ijk\ell} = b_i b_j c_{k\ell}$

$\mathcal{A} = \mathbf{C}^{\otimes 2} = \mathbf{C} \otimes \mathbf{C} \in \mathbb{R}^{n \times n \times n \times n} \iff a_{ijk\ell} = c_{ij} c_{k\ell}$

Symmetrization



$$(\mathrm{sym}(\boldsymbol{\mathcal{A}}))_{ijk} = \frac{1}{6}(a_{ijk} + a_{ikj} + a_{jik} + a_{jki} + a_{kij} + a_{kji})$$

Lemma (Hackbusch, 2019)

$$\langle \mathrm{sym}(\boldsymbol{\mathcal{A}}), \mathbf{b}^{\otimes d} \rangle = \langle \boldsymbol{\mathcal{A}}, \mathbf{b}^{\otimes d} \rangle$$

BONUS

Symmetric Tensors Correspond to Polynomials

MathSci.ai

$\mathcal{A} \in \mathbb{R}^{n \times n \times n}$

$$\Phi[\mathcal{A}](z_1, \dots, z_n) \equiv \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk} z_i z_j z_k = \langle \mathcal{A}, \mathbf{z}^{\otimes 3} \rangle$$

Proposition (Kileel-K-Pereira 2022)

$$\Phi[\mathbf{a}](z_1, \dots, z_n) = \mathbf{a}^\intercal \mathbf{z}$$

$$\Phi[\mathbf{A}](z_1, \dots, z_n) = \mathbf{z}^\intercal \mathbf{A} \mathbf{z}$$

$$\Phi[\mathcal{A}](z_1, \dots, z_n) = \Phi[\mathrm{sym}(\mathcal{A})](z_1, \dots, z_n)$$

$$\Phi[\mathcal{A} \otimes \mathcal{B}](z_1, \dots, z_n) = \Phi[\mathcal{A}](z_1, \dots, z_n) \cdot \Phi[\mathcal{B}](z_1, \dots, z_n)$$

Binomial Theorem for Tensors

(Kileel-K-Pereira 2022)

$$(\mathbf{a} + \mathbf{b})^{\otimes d} = \sum_{k=0}^{d} \binom{d}{k} \mathrm{sym}\left(\mathbf{a}^{\otimes k} \otimes \mathbf{b}^{\otimes d-k}\right)$$

MathSci.ai

Method of Moments for Gaussian Mixture Models – General Scenario

Gaussian Mixture Model: General Case

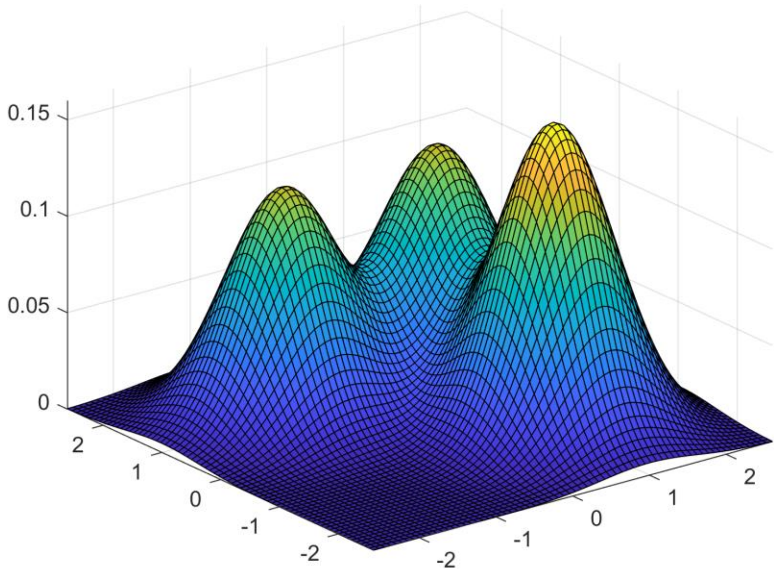
MathSci.ai

$$X \sim \sum_{j=1}^{m} \lambda_j \, \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

$$\min_{\theta} f(\theta) \equiv \left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} - \mathcal{M}^{(d)} \right\|^2$$

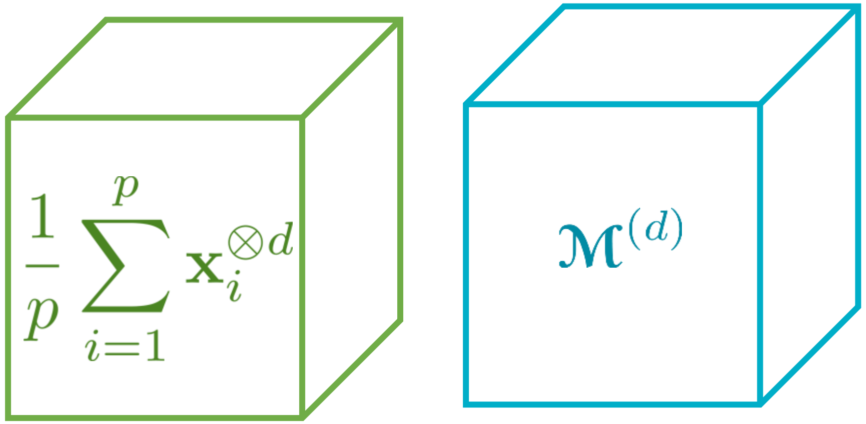
Optimization Parameters

$$\theta = \{ \lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \}_{j=1}^{m}$$



empirical moment tensor

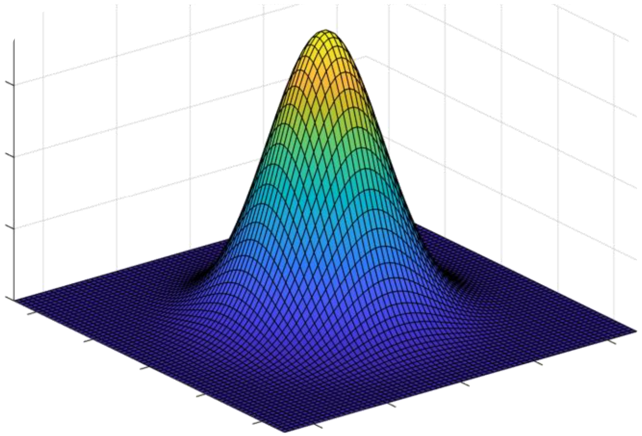
model moment tensor



Problem: Can we *explicitly* characterize the model moment tensor in terms of the parameters?

Gaussian Model Moment

Let $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathcal{M}^{(d)} = \mathbb{E}(X^{\otimes d})$. Then



$$\mathcal{M}^{(1)} = \boldsymbol{\mu}$$
$$\mathcal{M}^{(2)} = \boldsymbol{\mu}^{\otimes 2} + \boldsymbol{\Sigma}$$
$$\mathcal{M}^{(3)} = \boldsymbol{\mu}^{\otimes 3} + 3\,\mathrm{sym}(\boldsymbol{\mu} \otimes \boldsymbol{\Sigma})$$
$$\mathcal{M}^{(4)} = \boldsymbol{\mu}^{\otimes 4} + 6\,\mathrm{sym}(\boldsymbol{\mu}^{\otimes 2} \otimes \boldsymbol{\Sigma}) + 3\,\mathrm{sym}(\boldsymbol{\Sigma}^{\otimes 2})$$

Theorem (Kileel-K-Pereira 2022)

$$\mathcal{M}^{(d)} = \sum_{k=0}^{\lfloor d/2 \rfloor} \binom{d}{2k} \frac{2k!}{k!2^k}\,\mathrm{sym}\left(\boldsymbol{\mu}^{\otimes d-2k} \otimes \boldsymbol{\Sigma}^{\otimes k}\right)$$

Proof techniques

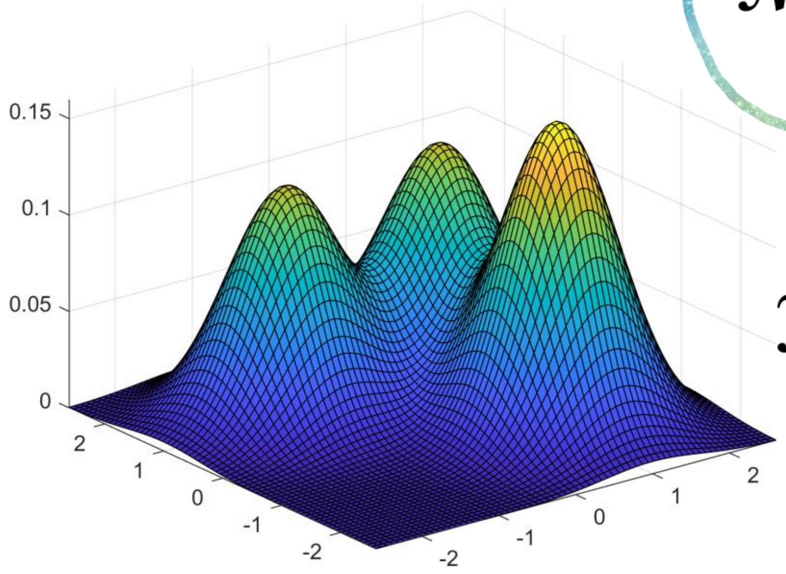
- Equivalence of symmetric tensors and polynomials
- Marginals of multivariate Gaussians are univariate Gaussian
- Binomial theorem

GMM Model Moment

<u>Theorem</u> (Kileel-K-Pereira 2022)

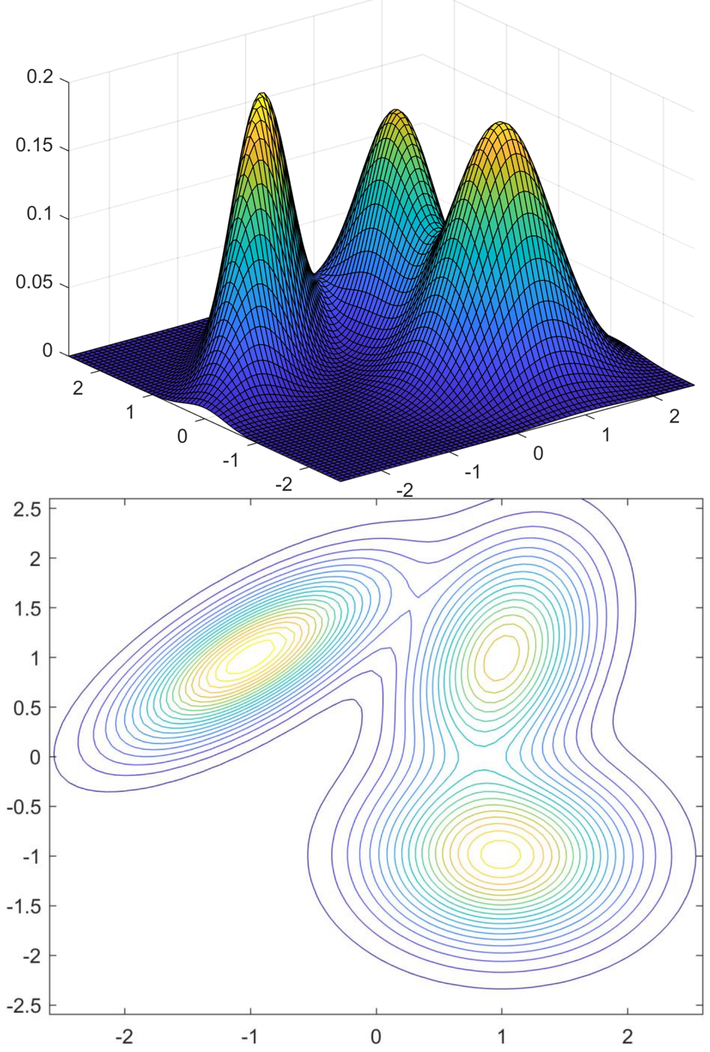
Let $X \sim \sum_{j=1}^m \lambda_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $\mathfrak{M}^{(d)} \equiv \mathbb{E}(X^{\otimes d})$.
Then

$$\mathfrak{M}^{(d)} = \sum_{j=1}^{m} \lambda_j \mathfrak{M}_j^{(d)}$$

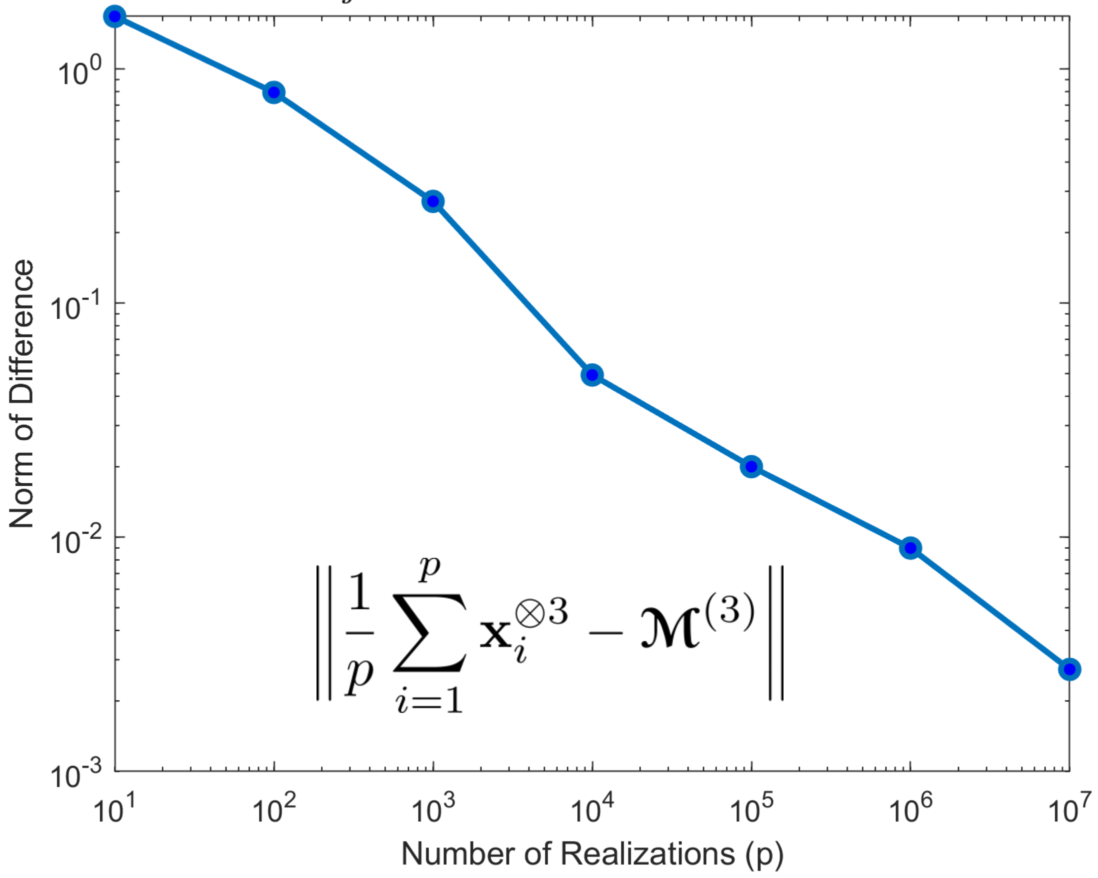


$$\mathfrak{M}_j^{(d)} = \sum_{k=0}^{\lfloor d/2 \rfloor} \binom{d}{2k} \frac{2k!}{k!2^k} \operatorname{sym}\left(\boldsymbol{\mu}_j^{\otimes d-2k} \otimes \boldsymbol{\Sigma}_j^{\otimes k} \right)$$

Empirical versus GMM Model Moment



$$\mathcal{M}^{(3)} = \sum_{j=1}^{3} \lambda_j \left(\boldsymbol{\mu}_j^{\otimes 3} + 3\,\mathrm{sym}(\boldsymbol{\mu}_j \otimes \boldsymbol{\Sigma}_j) \right)$$



Previous Approach was Biased

New Method: Exact

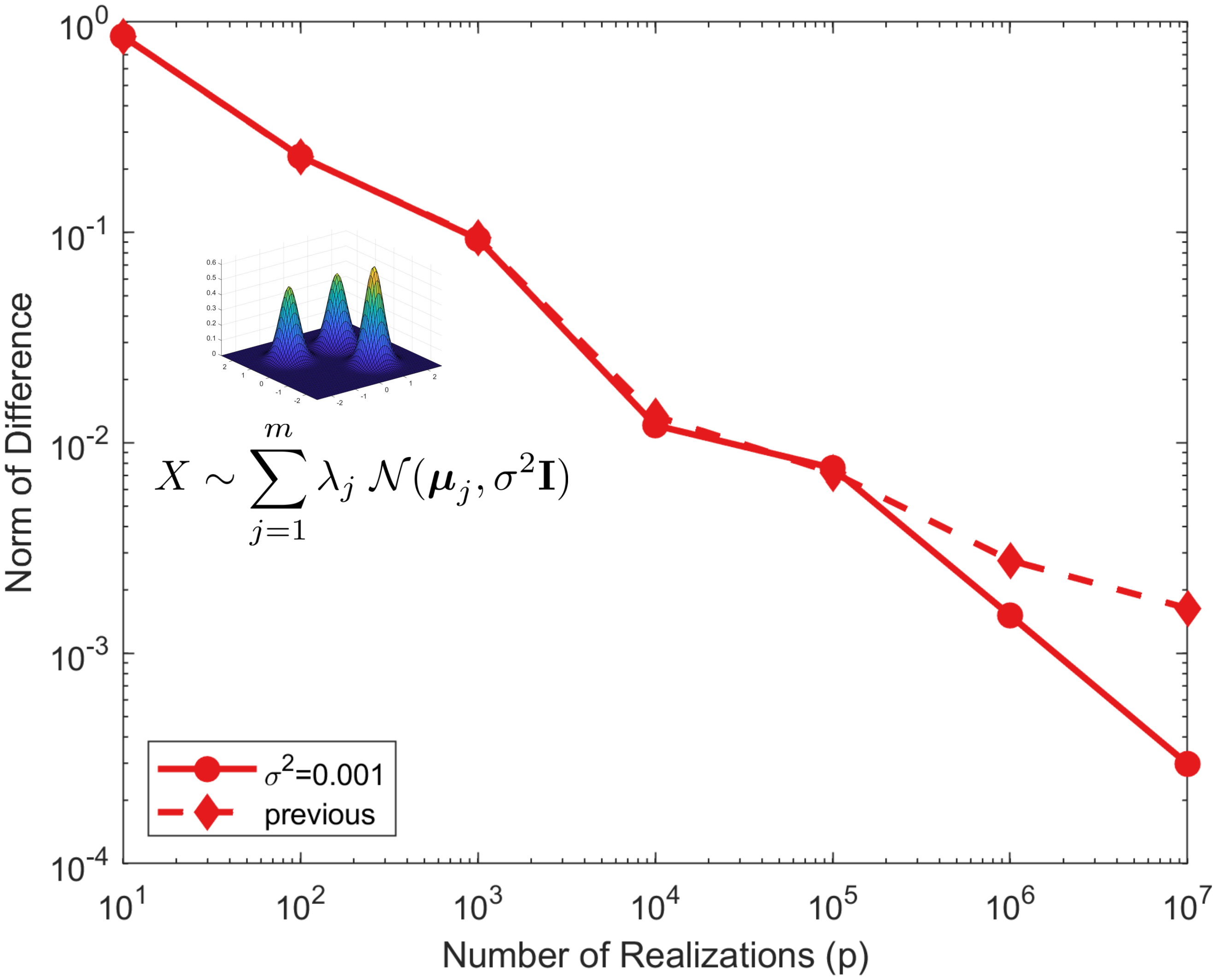
$$\mathcal{M}_j^{(3)} = \boldsymbol{\mu}_j^{\otimes 3} + 3\,\mathrm{sym}(\boldsymbol{\mu}_j \otimes \boldsymbol{\Sigma}_j)$$

$$\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes 3} - \sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(3)} \right\|$$

Previous Method: Approximate

$$\widetilde{\mathcal{M}}_j^{(3)} = \boldsymbol{\mu}_j^{\otimes 3}$$

$$\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes 3} - \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes 3} \right\|$$



Previous Approach was Biased

New Method: Exact

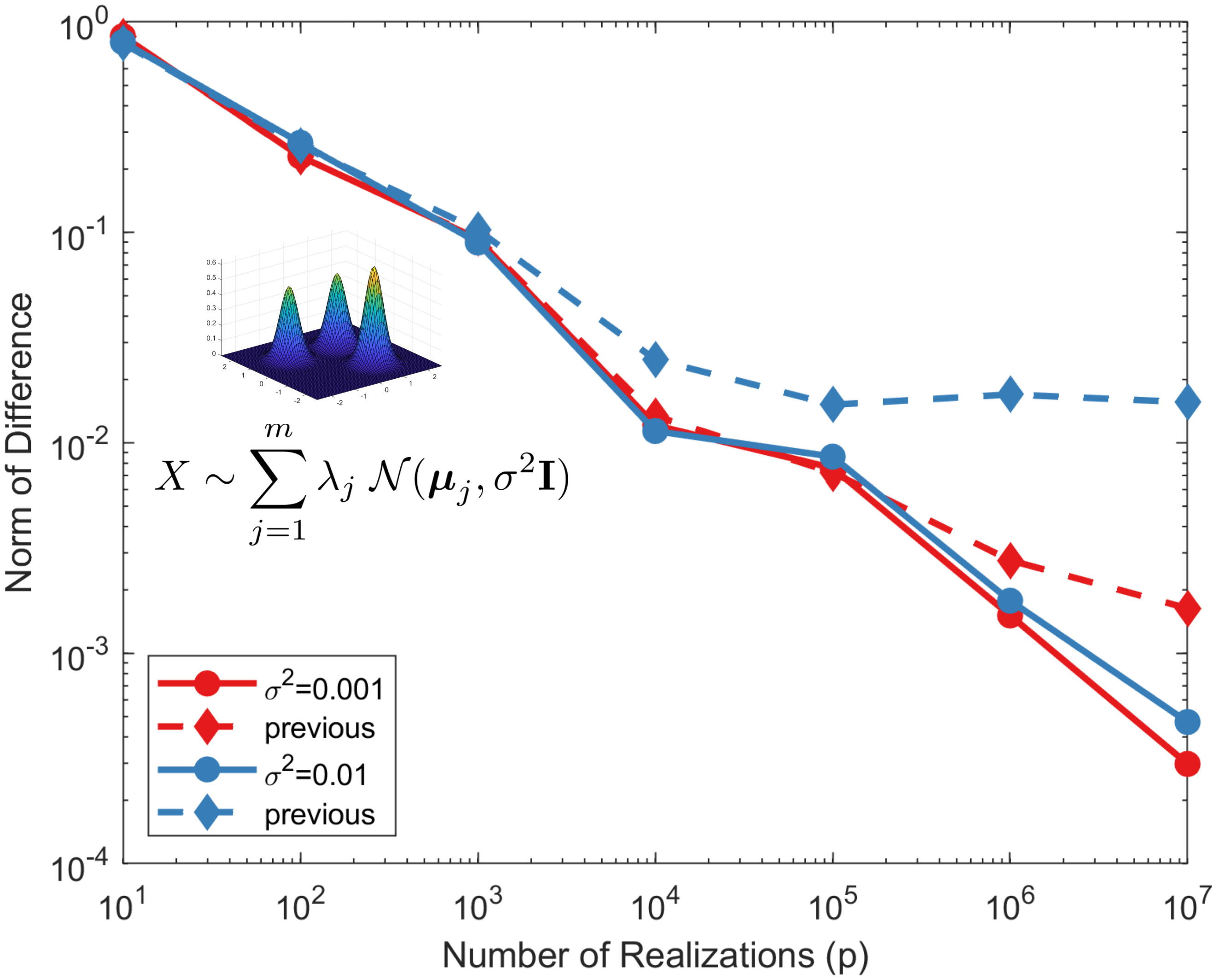
$$\mathfrak{M}_j^{(3)} = \boldsymbol{\mu}_j^{\otimes 3} + 3\,\mathrm{sym}(\boldsymbol{\mu}_j \otimes \boldsymbol{\Sigma}_j)$$

$$\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes 3} - \sum_{j=1}^{m} \lambda_j \mathfrak{M}_j^{(3)} \right\|$$

Previous Method: Approximate

$$\widetilde{\mathfrak{M}}_j^{(3)} = \boldsymbol{\mu}_j^{\otimes 3}$$

$$\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes 3} - \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes 3} \right\|$$



Previous Approach was Biased

New Method: Exact

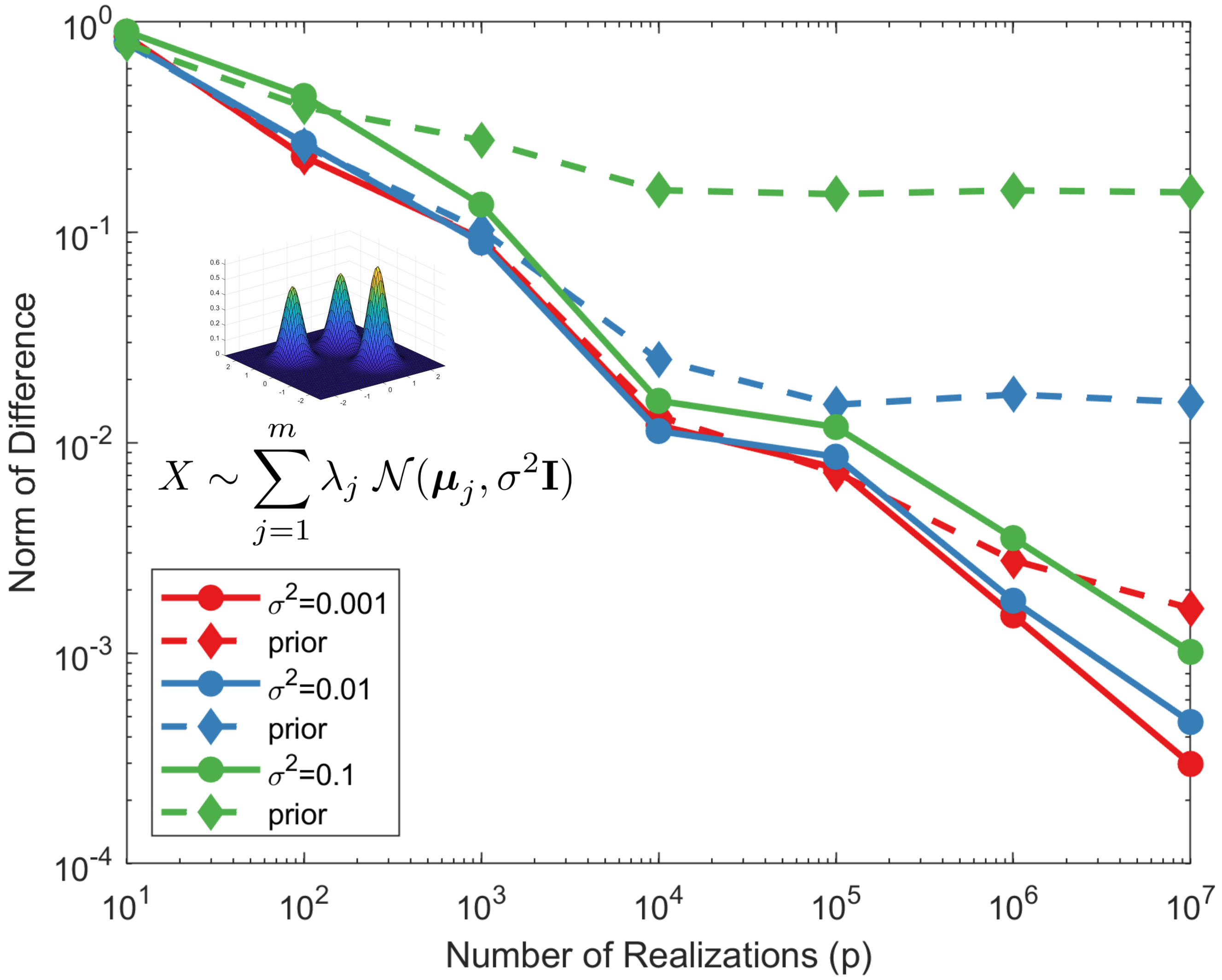
$$\mathcal{M}_j^{(3)} = \boldsymbol{\mu}_j^{\otimes 3} + 3\,\mathrm{sym}(\boldsymbol{\mu}_j \otimes \boldsymbol{\Sigma}_j)$$

$$\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes 3} - \sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(3)} \right\|$$

Previous Method: Approximate

$$\widetilde{\mathcal{M}}_j^{(3)} = \boldsymbol{\mu}_j^{\otimes 3}$$

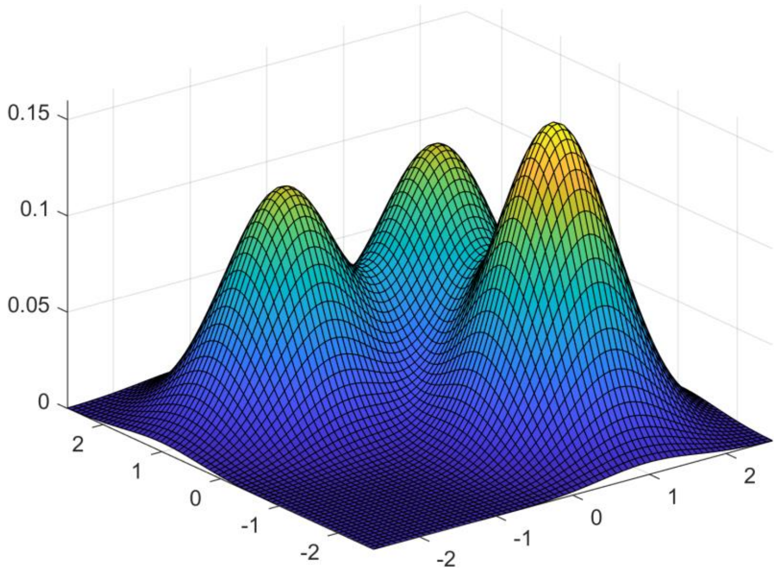
$$\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes 3} - \sum_{j=1}^{m} \lambda_j \boldsymbol{\mu}_j^{\otimes 3} \right\|$$



Optimization Formulation

MathSci.ai

$$X \sim \sum_{j=1}^{m} \lambda_j \, \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$



$$\min_{\theta} f(\theta) \equiv \left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} - \mathcal{M}^{(d)} \right\|^2$$

empirical moment tensor

$$\frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d}$$

model moment tensor

$$\sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(d)}$$

Optimization Parameters

$$\theta = \{ \lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \}_{j=1}^{m}$$

Problem: Can we *explicitly* characterize the model moment tensor in terms of the parameters?

Yes!

$$\mathcal{M}_j^{(d)} = \sum_{k=0}^{\lfloor d/2 \rfloor} \binom{d}{2k} \frac{2k!}{k! 2^k} \, \mathrm{sym}\left(\boldsymbol{\mu}_j^{\otimes d-2k} \otimes \boldsymbol{\Sigma}_j^{\otimes k} \right)$$

Optimization Formulation

MathSci.ai

Optimization Parameters

$$\theta = \{\, \lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \,\}_{j=1}^{m}$$

$$\min_{\theta} f(\theta) \equiv \left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} - \sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(d)} \right\|^2$$

empirical moment tensor

$$\frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d}$$

model moment tensor

$$\sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(d)}$$

$$\mathcal{M}_j^{(d)} = \sum_{k=0}^{\lfloor d/2 \rfloor} \binom{d}{2k} \frac{2k!}{k! 2^k} \, \mathrm{sym}\left(\boldsymbol{\mu}_j^{\otimes d-2k} \otimes \boldsymbol{\Sigma}_j^{\otimes k} \right)$$

Optimization Formulation as Inner Products

MathSci.ai

$$\min_{\theta} f(\theta) \equiv \left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} - \sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(d)} \right\|^2$$

$$f(\theta) = \cancel{\left\| \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d} \right\|^2} + \left\| \sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(d)} \right\|^2 - 2 \left\langle \frac{1}{p} \sum_{i=1}^{p} \mathbf{x}_i^{\otimes d}, \sum_{j=1}^{m} \lambda_j \mathcal{M}_j^{(d)} \right\rangle$$

constant

$$f(\theta) = C + \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j \underbrace{\left\langle \mathcal{M}_i^{(d)}, \mathcal{M}_j^{(d)} \right\rangle}_{\text{dot product of 2 moments}} - \frac{2}{p} \sum_{i=1}^{p} \sum_{j=1}^{m} \lambda_j \underbrace{\left\langle \mathbf{x}_i^{\otimes d}, \mathcal{M}_j^{(d)} \right\rangle}_{\text{dot product of moment + vector}}$$

See forthcoming arXiv posting for full details!!

Example Calculation: $d = 3$

MathSci.ai

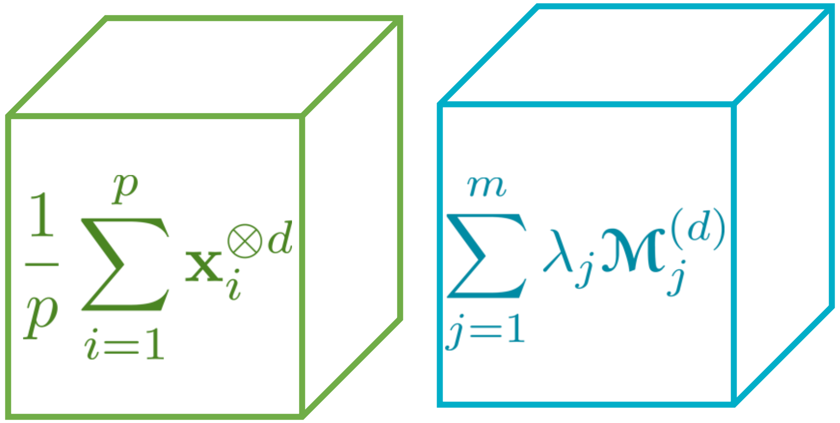
$$f(\theta) = C + \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \left\langle \mathcal{M}_i^{(d)}, \mathcal{M}_j^{(d)} \right\rangle - \frac{2}{p} \sum_{i=1}^{p}\sum_{j=1}^{m} \lambda_j \left\langle \mathbf{x}_i^{\otimes d}, \mathcal{M}_j^{(d)} \right\rangle$$

$$\mathcal{M}_j^{(3)} = \boldsymbol{\mu}_j^{\otimes 3} + 3\,\mathrm{sym}(\boldsymbol{\mu}_j \otimes \boldsymbol{\Sigma}_j)$$

$$\begin{aligned}
\langle \mathbf{x}_i^{\otimes 3}, \mathcal{M}_j^{(3)} \rangle &= \langle \mathbf{x}_i^{\otimes 3}, \boldsymbol{\mu}_j^{\otimes 3} \rangle + 3 \langle \mathbf{x}_i^{\otimes 3}, \mathrm{sym}(\boldsymbol{\mu}_j \otimes \boldsymbol{\Sigma}_j) \rangle \\
&= (\mathbf{x}_i^{\intercal} \boldsymbol{\mu}_j)^3 + 3 \langle \mathbf{x}_i^{\otimes 3}, \boldsymbol{\mu}_j \otimes \boldsymbol{\Sigma}_j \rangle \\
&= (\mathbf{x}_i^{\intercal} \boldsymbol{\mu}_j)^3 + 3 (\mathbf{x}_i^{\intercal} \boldsymbol{\mu}_j)(\mathbf{x}_i^{\intercal} \boldsymbol{\Sigma}_j \mathbf{x}_i)
\end{aligned}$$

Theory for computing inner product of two GMM moment tenors or empirical and GMM moment tensor – see forthcoming arXiv paper for details!

Casting as Optimization Problem



$$\min_{\theta} f(\theta) = C + \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \left\langle \mathcal{M}_i^{(d)}, \mathcal{M}_j^{(d)} \right\rangle - \frac{2}{p}\sum_{i=1}^{p}\sum_{j=1}^{m} \lambda_j \left\langle \mathbf{x}_i^{\otimes d}, \mathcal{M}_j^{(d)} \right\rangle$$

$$\theta = \{\, \lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \,\}_{j=1}^{m}$$

- Never need to form empirical or model moments explicitly, overcoming curse of dimensionality
- Function can be calculated using simple calculations, total work $\mathcal{O}(m^2 n + pmn^2 + m^2 n^3)$ per iteration and $\mathcal{O}(mn + pn)$ storage
- Gradients can be calculated as well
- Easy stochastic function and gradient if number of samples (p) is large, as before
- *Issue:* Inherent scaling problem

Non-uniqueness problem

$$X \sim 0.5\, \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.5\, \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

$$Y \sim 0.25\, \mathcal{N}\left((\tfrac{1}{2})^{1/3} \boldsymbol{\mu}_1, (\tfrac{1}{2})^{2/3} \boldsymbol{\Sigma}_1 \right) + 0.75\, \mathcal{N}\left((\tfrac{3}{2})^{1/3} \boldsymbol{\mu}_2, (\tfrac{3}{2})^{2/3} \boldsymbol{\Sigma}_2 \right)$$

$$\mathbb{E}(X^{\otimes 3}) = \mathbb{E}(Y^{\otimes 3})$$

Non-uniqueness problem

MathSci.ai

$$X \sim \sum_{j=1}^{m} \lambda_j \, \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \qquad Y \sim \sum_{j=1}^{m} \lambda_j \gamma_j^{-d} \, \mathcal{N}(\gamma_j \boldsymbol{\mu}_j, \gamma_j^2 \boldsymbol{\Sigma}_j)$$

Problem: Moments of order d are the same!

FIX: Append a constant c to the end of every observation vector, creating vectors of dimension $n + 1$

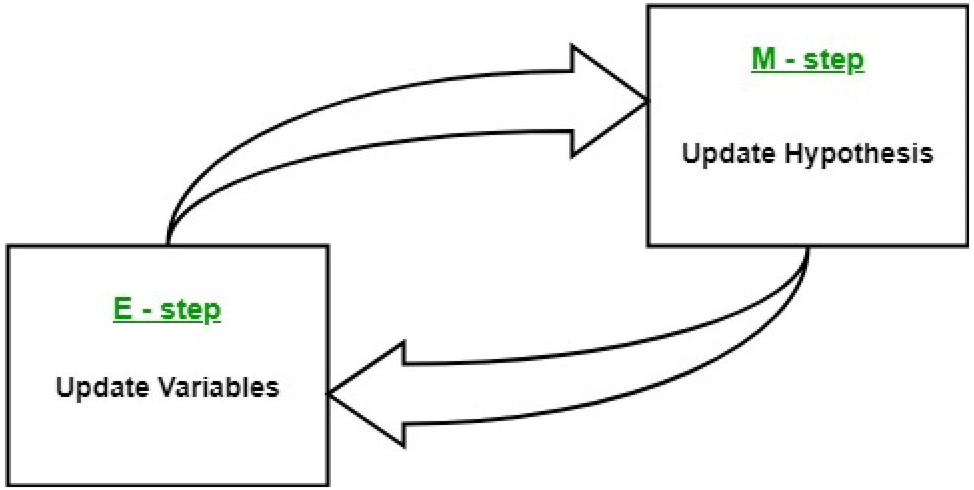
RESULT: *Implicitly,* a weighted combination of *all the moments* from 1 to d. This means we include all moments up to order d in the optimization.

See forthcoming arXiv posting for full details!!

MathSci.ai

Numerical Results and Conclusions

State of the Art: Expectation Maximization (EM)



EM Algorithm in a Nutshell
Make initial guesses for parameters
Repeat until log-likelihood converges:

1. Compute membership weights for each datapoint
2. Update the component parameters using the membership weights

EM is State of the Art

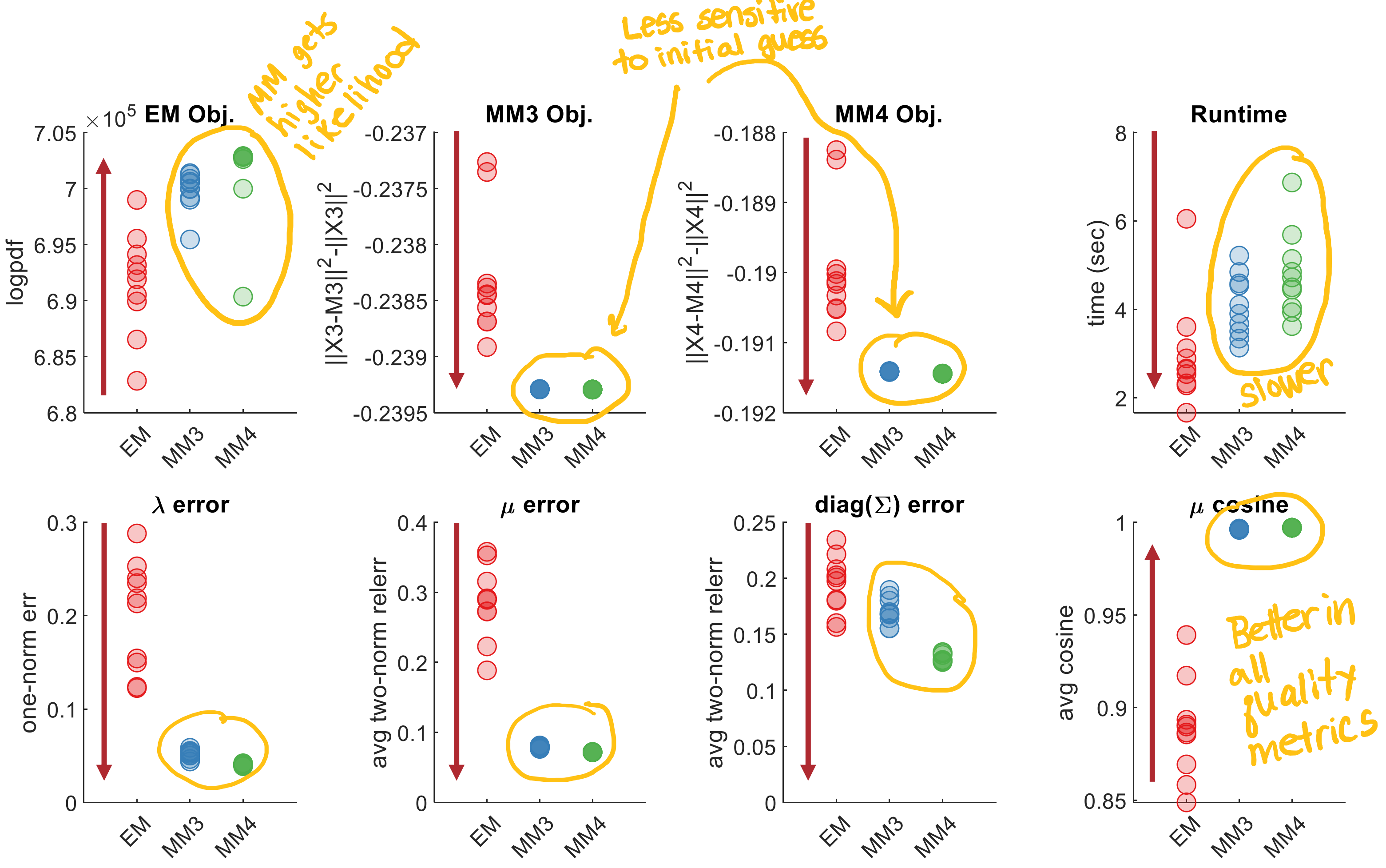
- Inexpensive
- Relatively easy to implement
- Optimizing a different cost function
- Sensitive to initialization
- Sensitive to overlapping Gaussians

MoM has theoretical advantages but has not been used much in practice previously because of its great expense

See, e.g., Xu and Jordan (1996) for discussion of its robustness

Method of Moments can beat EM

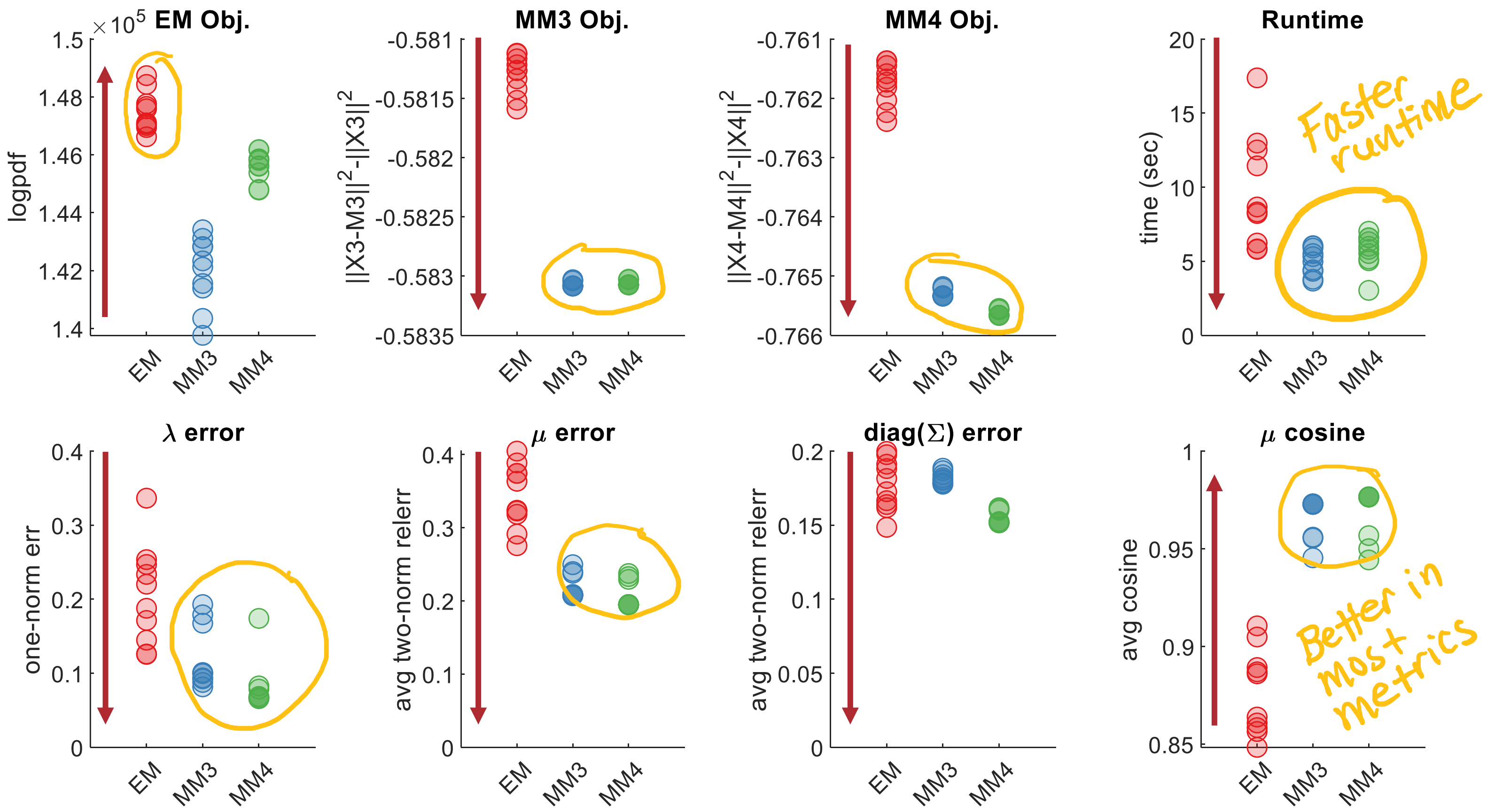
- Randomly-generated problem with overlapping Gaussians
  - Diagonal covariances
  - Dimensionality: $n = 100$
  - Number of Gaussians: $m = 20$
  - Observations: $p = 8000$
- Compared three methods
  - EM: Expectation Maximization
  - MM3: Method of Moments, $d = 3$
  - MM4: Method of Moments, $d = 4$
- 10 runs each with different initial guesses



Method of Moments can beat EM

- Randomly-generated problem with overlapping Gaussians
  - Diagonal covariances
  - Dimensionality: $n = 100$
  - Number of Gaussians: $m = 20$
  - Observations: $p = 8000$
- Compared three methods
  - EM: Expectation Maximization
  - MM3: Method of Moments, $d = 3$
  - MM4: Method of Moments, $d = 4$
- 10 runs each with different initial guesses

Same setup as previous slide except higher noise

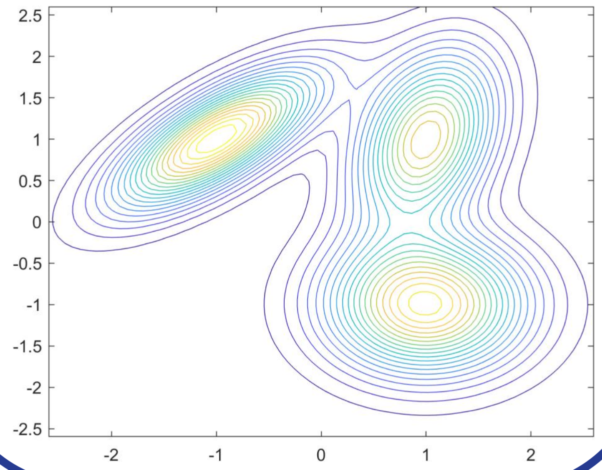


Related Works

MathSci.ai

Key Differences in Our Work

- Novel tensor formulation of Gaussian moments
- No spherical or axis-aligned covariance assumptions
- Computationally efficient, no exponential dependence on d

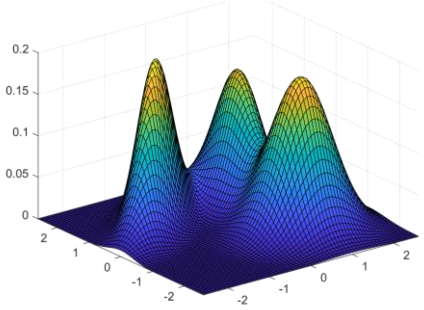


- Many theoretical advantages to method of moments and connections to tensors
  - Hsu and Kakade (2013) – diagonal covariance, $d \leq 3$
  - Ge, Huang, Kakade (2015) – vectorized covariances, loses symmetries
  - Bakshi, Diakonikolas, Jia, Kane, Kothari, and Vempala (2020) – robust learning using tensor decomposition
  - Khouga, Mattei, Mourrian (2021) – GMM identifiability
- Computational approaches (limited handling of covariances)
  - Anandkumar, Ge, Hsu, Kakade (2014) and Anandkumar, Ge, Hsu, Kakade, Telegarsky (2014) – orthogonal symmetric tensor decomposition
  - Sherman & K., 2020 – (general) symmetric tensor decomposition, emphasis of implicit computation to avoid curse of dimensionality
- Inner products of moment tensors
  - Muandet, Fukumizu, Dinuzzo, Schölkopf (2012) – up to 3rd order

Take-aways and Future Work

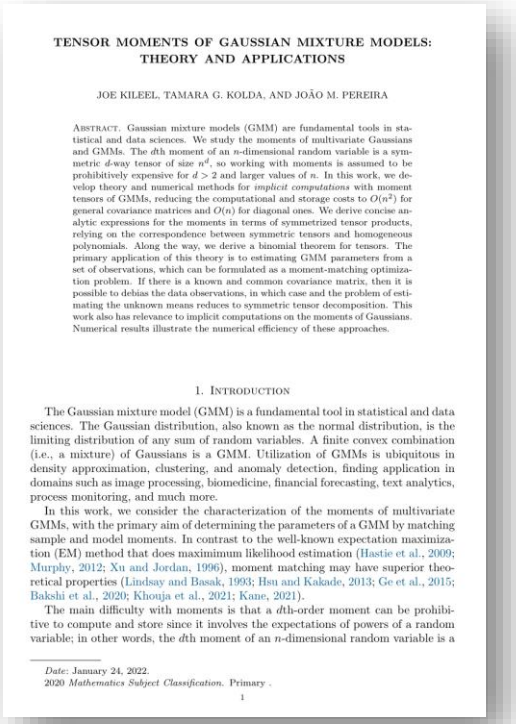
- Our focus: Method of moments for Gaussian mixture models (GMMs)
- Key results
  - Formulation of GMM moment in terms of tensor outer products

$$\boxed{\mathcal{M}^{(d)} = \sum_{j=1}^{m} \sum_{k=0}^{\lfloor d/2 \rfloor} \binom{d}{2k} \frac{2k!}{k!2^k} \lambda_j \, \mathrm{sym}\left(\boldsymbol{\mu}_j^{\otimes d-2k} \otimes \boldsymbol{\Sigma}_j^{\otimes k} \right)}$$



  - Efficient computation and storage, avoiding exponential dependence on moment order
  - Novel approach to scaling ambiguity using augmentation
  - Amenable to stochastic formulations
  - Plus…dot product of moment tensors in terms of Bell polynomials, avoiding exponential dependence on moment order
  - Plus…modifying empirical moment tensor to "remove" Gaussian noise
- Future work
  - Implementation details, especially for general Gaussians
  - Analysis of optimization landscape and comparison to that of max likelihood
  - Bounding number of samples required for accurate estimation
  - Application studies

See forthcoming arXiv posting for full details!!



TENSOR MOMENTS OF GAUSSIAN MIXTURE MODELS: THEORY AND APPLICATIONS

JOE KILEEL, TAMARA G. KOLDA, AND JOÃO M. PEREIRA

Abstract. Gaussian mixture models (GMM) are fundamental tools in statistical and data sciences. We study the moments of multivariate Gaussians and GMMs. The dth moment of an n-dimensional random variable is a symmetric d-way tensor of size n^d, so working with moments is assumed to be prohibitively expensive for $d > 2$ and larger values of n. In this work, we develop theory and numerical methods for *implicit computations* with moment tensors of GMMs, reducing the computational and storage costs to $O(n^2)$ for general covariance matrices and $O(n)$ for diagonal ones. We derive concise analytic expressions for the moments in terms of symmetrized tensor products, relying on the correspondence between symmetric tensors and homogeneous polynomials. Along the way, we derive a binomial theorem for tensors. The primary application of this theory is to estimating GMM parameters from a set of observations, which can be formulated as a moment-matching optimization problem. If there is a known and common covariance matrix, then it is possible to debias the data observations, in which case and the problem of estimating the unknown means reduces to symmetric tensor decomposition. This work also has relevance to implicit computations on the moments of Gaussians. Numerical results illustrate the numerical efficiency of these approaches.

1. Introduction

The Gaussian mixture model (GMM) is a fundamental tool in statistical and data sciences. The Gaussian distribution, also known as the normal distribution, is the limiting distribution of any sum of random variables. A finite convex combination (i.e., a mixture) of Gaussians is a GMM. Utilization of GMMs is ubiquitous in density approximation, clustering, and anomaly detection, finding application in domains such as image processing, biomedicine, financial forecasting, text analytics, process monitoring, and much more.

In this work, we consider the characterization of the moments of multivariate GMMs, with the primary aim of determining the parameters of a GMM by matching sample and model moments. In contrast to the well-known expectation maximization (EM) method that does maximimum likelihood estimation (Hastie et al., 2009; Murphy, 2012; Xu and Jordan, 1996), moment matching may have superior theoretical properties (Lindsay and Basak, 1993; Hsu and Kakade, 2013; Ge et al., 2015; Bakshi et al., 2020; Khouja et al., 2021; Kane, 2021).

The main difficulty with moments is that a dth-order moment can be prohibitive to compute and store since it involves the expectations of powers of a random variable; in other words, the dth moment of an n-dimensional random variable is a

Date: January 24, 2022.
2020 *Mathematics Subject Classification*. Primary .

1

J. Kileel, T. Kolda, and J. M. Pereira. Tensor Moments of Gaussian Mixture Models: Theory and Applications