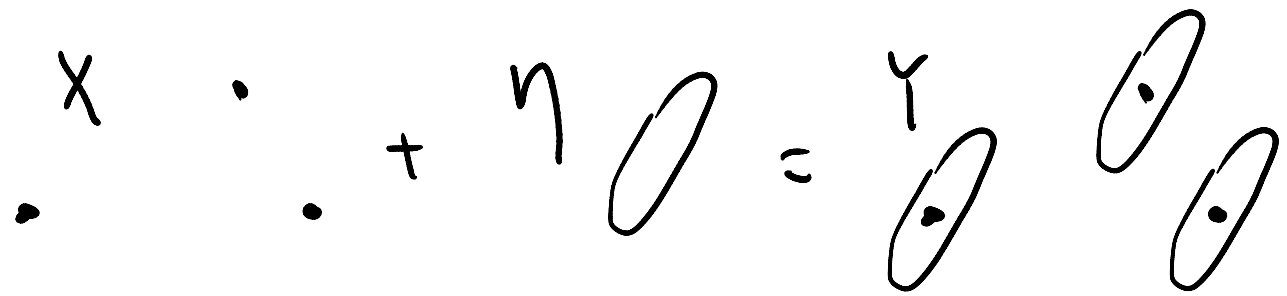
The More, the Merrier: the Blessing of Dimensionality for Learning Large Gaussian Mixtures

Joseph Anderson (OSU)
Mikhail Belkin (OSU)
Navin Goyal (MSR)
Luis Rademacher (OSU)
James Voss (OSU)




A (simplified) question

- Given samples from random vector $Y = X + \eta$ where X is uniform in a discrete set in R^n and $\eta \sim N(0, I)$, estimate the support of X.
- Stability of recovery? From moments?
- How many samples are needed?
- Efficient algorithm?



Computational learning

- **Efficient** high dimensional inference:
  - Given samples from a distribution, estimate something about the distribution.
- Learn = estimate.

Questions

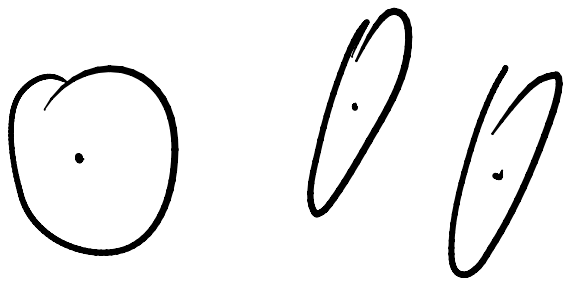
- Identifiability in parameter estimation (uniqueness): different parameters imply different distributions.
- Robust identifiability (stability and sample complexity): far in parameter space $\Rightarrow$ far in distribution. Implies identifiability.
- Efficient estimation (computational complexity). Implies robust identifiability.
- Interested in polynomial v/s exponential dependencies. “Far” means 1/poly(.) while “close” means 1/exp(.). Efficient means polynomial time.

“Easy” Example

- n-dimensional Gaussians: family of distributions parameterized by mean and covariance matrix. Clearly identifiable.
- Robustly identifiable: If two Gaussians have similar parameters, then they are close as distributions (say, via K-L divergence and Pinsker’s inequality).
- Efficient estimation: empirical mean and covariance are close to true values with high probability given $poly(n)$ samples.

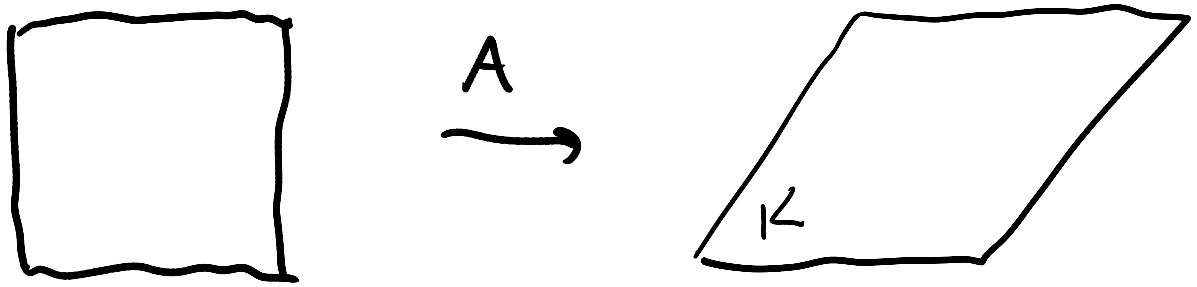
Harder Example

- (Finite) Gaussian mixture:
  Given samples from n-dimensional density of the form $f = \sum_{i=1}^{k} w_i f_i$, where f_i is $N(\mu_i, \Sigma_i)$, $w_i > 0$, $\sum w_i = 1$, estimate w_i, μ_i, Σ_i.
- Identifiability [Teicher '61]: Yes, up to permutation and mixtures having two identical components.
- Robust identifiability and complexity? Later.



Example: Learn a parallelepiped

- Given uniformly random points from a linear transformation of a hypercube, estimate the linear transformation (up to inherent ambiguities).

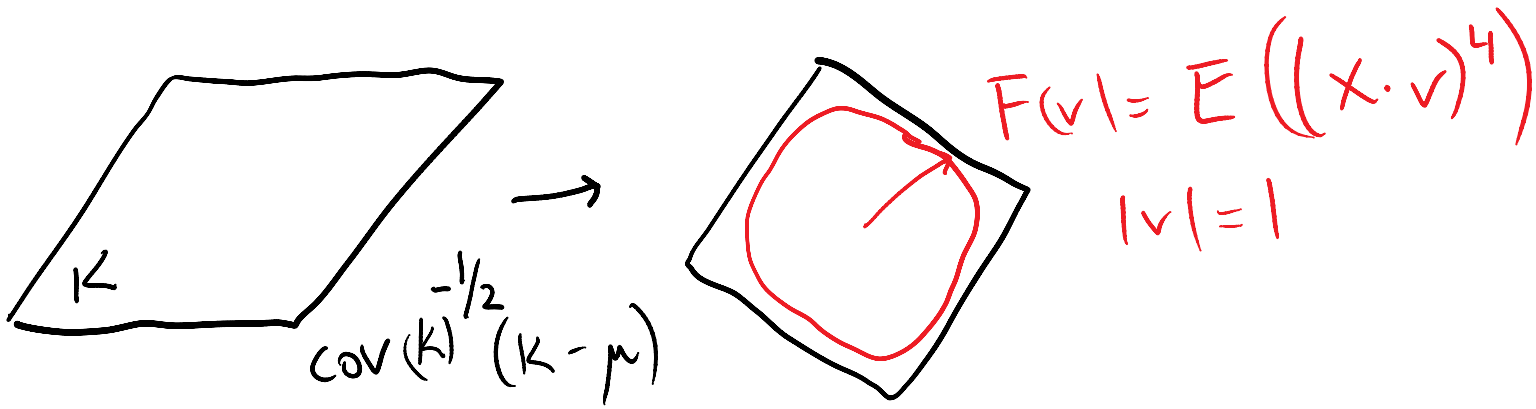


Example: Learn a parallelepiped

- [Frieze Jerrum Kannan] [Nguyen Regev] Learn a parallelepiped in polynomial time.
- [FJK] Actually, Independent Component Analysis (ICA): Given d-dim samples from Y given by $Y = AX + b$, can estimate A, b, for X with unknown d-dim. distribution having *independent components*.

How to learn a parallelepiped? [FJK]

- By estimating mean and covariance, can assume it is a rotated cube centered at 0
- To estimate rotation: Enumerate all local minima of directional 4^{th} moment on unit sphere. Normals to facets are a complete set of local minima.



Actually…

- For X isotropic and having independent coordinates, [FJK] show:

$$E((v \cdot X)^4) = 3 + \sum v_i^4 \left(E\left(X_i^4\right) - 3\right)$$

- Why "-3"? What if $E\left(X_i^4\right) = 3$? (4^{th} moment would be constant as v varies, so local optima give nothing).
- Let $\kappa_4(Z) = E(Z^4) - 3E(Z^2)^2$. Then

$$\kappa_4(v \cdot X) = \sum v_i^4 \, \kappa_4(X_i)$$

“Hidden” generalized characteristic function

- Let $\psi_X(t) = \log E(e^{t \cdot X})$. Let $\kappa_j(X)$ be given by the power series expansion of ψ:

$$\psi_X(t) = \sum_{j=1}^{\infty} \kappa_j(X) \frac{t^j}{j!}$$

- $\kappa_j(X)$: cumulants
- $\psi_X(t)$: cumulant generating function
- κ_1 =mean, κ_2 =variance, $\kappa_4(X) = E(X^4) - 3E(X^2)^2$ (equating coefficient of CGF and MGF)
- $\kappa_j(aX) = a^j \kappa_j(X)$
- $\kappa_j(X + Y) = \kappa_j(X) + \kappa_j(Y)$ for independent X, Y
- If a real valued functional is continuous and additive in the space of random variables with moments, then it is a linear combination of cumulants.

Yeredor's idea for ICA even if $E\left(X_i^4\right) = 3$

- If $Y = AX$ where X has independent coordinates, we have
$$M(t) \coloneqq \psi_Y''(t) = A\psi_X''(A^T t)A^T$$
where $\psi_X''(A^T t)$ is a diagonal matrix.
- Then $M(t_1)M(t_2)^{-1} = ADA^{-1}$ for some diagonal matrix D and diagonalization of a sample estimate of $M(t_1)M(t_2)^{-1}$ (for random unit t_1, t_2) recovers A up to scale and permutation.

Another example: Simplex

- [Anderson Goyal R., “Efficient learning of simplices”] Efficient algorithm from first 3 moments.
  Idea: estimate first two moments to put the simplex in (approximate) isotropic position. Then maximize 3rd directional moment over sphere. Set of maxima = set of vertices.
- [Anandkumar, Foster, Hsu, Kakade, Liu] [Anandkumar, Ge, Hsu, Kakade, Telgarsky] Similar results, and for the more general Dirichlet distribution.

Polytopes with few vertices?

- Can one estimate a polytope with few vertices efficiently? (by S. Vempala)
- [Gravin Lasserre Pasechnik Robbins]
  To reconstruct a polytope with d vertices in R^n: Project onto a line. Recover projection of vertices using first $O(dn)$ moments. Repeat for $O(d)$ random lines.
  In principle, moment dn implies unstable and inefficient.
- Open identifiability question: Identifiability of (generic?) polytopes with $poly(n)$ (say, "n^2") vertices from first "100" moment tensors. Stability? (by N. Goyal)

A simpler problem

- Convex hull is difficult…
- What about the recovery of a discrete distribution via moments?
- Better motivated in practice: recover a discrete distribution with additive Gaussian noise. A special case of parameter estimation of a Gaussian mixture.

Robustness and efficiency in estimation of Gaussian mixtures

- Consider mixtures of k Gaussians in R^n.
  Given samples from n-dimensional density of the form, $f = \sum_{i=1}^{k} w_i f_i$, where f_i is $N(\mu_i, \Sigma_i)$, $w_i > 0$, $\sum w_i = 1$, estimate μ_i, w_i.
  For simplicity, focus on estimation of means and weights and assume convenient structure of covariance matrices (identical, spherical).
- Efficient algorithm implies low sample complexity and robust identifiability.
- Two kinds of results until recently:
  - With separation assumption:
    [Dasgupta] [Arora Kannan] [...]
    [Vempala Wang] If means are at distance $\sigma_{\max}\sqrt{k}$, can learn efficiently in n, k.
  - Without separation:
    [Belkin Sinha '10] [Moitra Valiant '10] Can learn in time poly(n) for any fixed k.
    Superexponential in k.
- [Moitra Valiant '10] Lower bound: exponential dependence in k necessary for $n = 1$. More precisely:
  Theorem: There exist two mixtures in R with q components at L^1 distance $\leq e^{-cq}$ and parameter distance $\geq c'/q$.
- Algorithms with and without separation are very different. What's between? Is there a unified view?

"Determined" ($k \leq n$) mixture with arbitrary separation

- [Hsu Kakade '12] Efficient learning of $k \leq n$ Gaussians in R^n. To quantify closeness to 1-dim hard case, complexity depends on $\sigma_k(A)$, the min singular value of the matrix of the means (complexity is $\text{poly}\left(\frac{1}{\sigma_k}\right)$).

"Underdetermined" ($k > n$) mixture

- What about $k > n$? What could play the role of σ_k?

Our Results

- For any fixed q, learn n^q Gaussians in R^n in smoothed polynomial time. More precisely:

  For any fixed q, for Gaussian mixtures in R^n with $k \leq O(n^q)$ (known) components and know identical covariance: can estimate means and weights in time polynomial in $1/s$, n, (and other obvious dependencies).
  - s is a **conditioning parameter** for the means. E.g. for n means, $\mathrm{s} = \sigma_{min}$ of matrix of means.
- Our algorithm reduces the problem to Independent Component Analysis.

Related work

- [Bhaskara Charikar Moitra Vijayaraghavan ‘13] Similar, simultaneous results: can learn mixtures of axis-aligned Gaussians with unknown covariance. Worse running time. Better smoothed analysis.

Conditioning parameter "s" for a mixture

- Given q, tensorize normalized means q times: Consider the set of tensors $\hat{\mu}_i^{\times q}$. Then

$$s = \sigma_k(\hat{\mu}_1^{\times q}, \dots, \hat{\mu}_k^{\times q})$$

Our Results

- Conditioning parameter s is at least inverse polynomial (in the smoothed analysis sense). E.g. for $k = \binom{n}{2}$ (and $q = 2$), (μ_i) any fixed set of k means in R^n and (E_i) independent vectors with entries $N(0, \sigma^2)$,

$$P\left(\sigma_k((\mu_i + E_i)^{\times 2}) \leq \frac{\sigma^2}{n^7}\right) = O\left(\frac{1}{n}\right)$$

Our Results

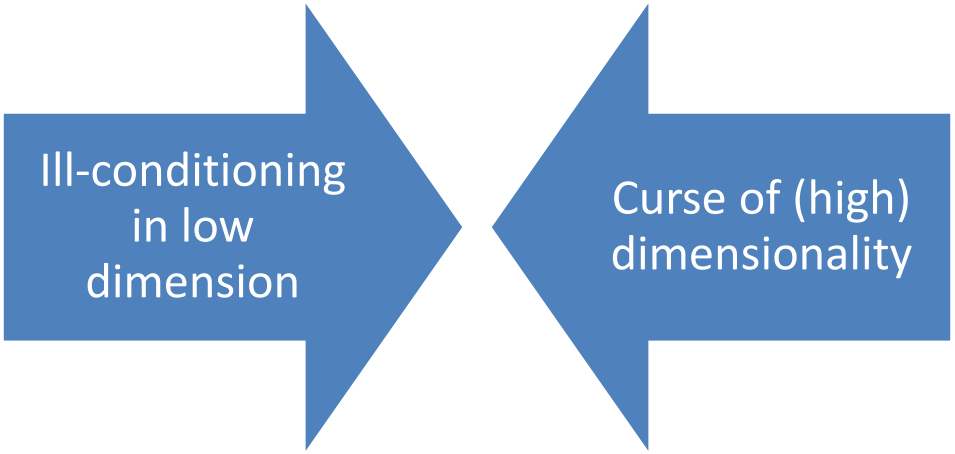
- Problem is **generically hard** in low dimension: Given k^2 random points from $[0,1]^n$, there exists two disjoint subsets A, B of k points and two mixtures M, N with means in A, B respectively and identical covariances so that

$$\|M - N\|_1 = d_{TV}(M, N) \leq e^{-k^{1/n}}$$

- This lower bound for GMM also applies to Independent Component Analysis (via the reduction).

Curse and blessing of dimensionality

- Concentration of mass can be good for estimation. Say, estimators concentrate around their means.
- It can be very bad: robust identifiability requires things to look different, distinguishable.
  - Powerful interpolation results for smooth functions in low dimension make the problem generically hard there.
- We want conditioning parameter s to be away from 0: need *anticoncentration.*
- Two forces – tradeoff:



Idea: Reduction to ICA

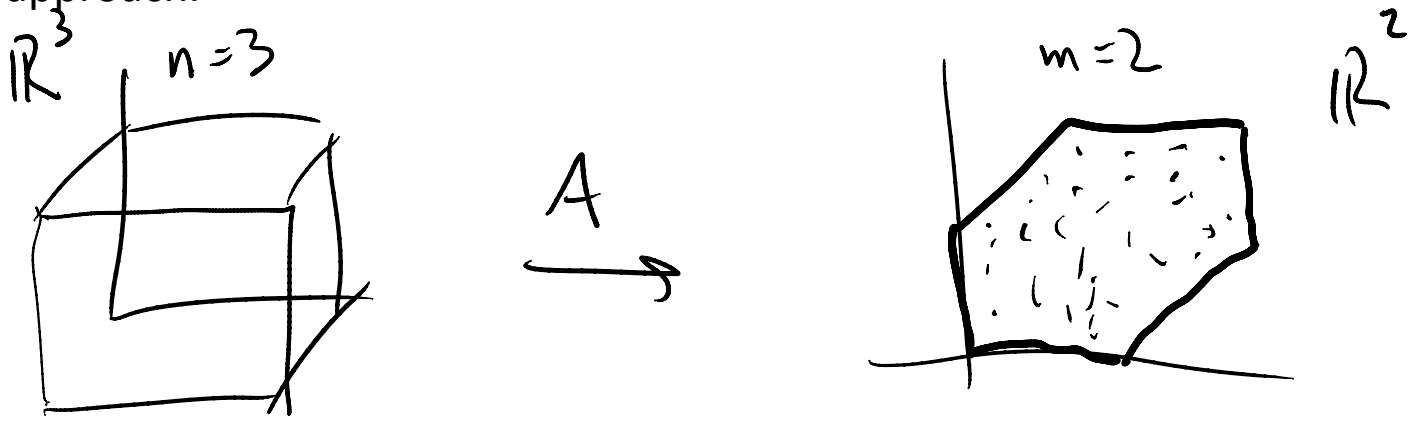
- Underdetermined ICA:
  Given samples from $X = AS$, where
  - S is random in R^n with independent coordinates,
  - A is m-by-n matrix with unit columns (think $m \leq n$).

  Estimate A (up to inherent ambiguities: sign and permutation of columns).

  "Recover a linear mapping of a random vector having independent coordinates".
- [Goyal Vempala Xiao '13]: Algorithm with provable sample and time guarantees. Also robust against additive independent Gaussian noise η:

  $$X = AS + \eta.$$

  It is a sophisticated generalization of Yeredor's cumulant generating function approach.

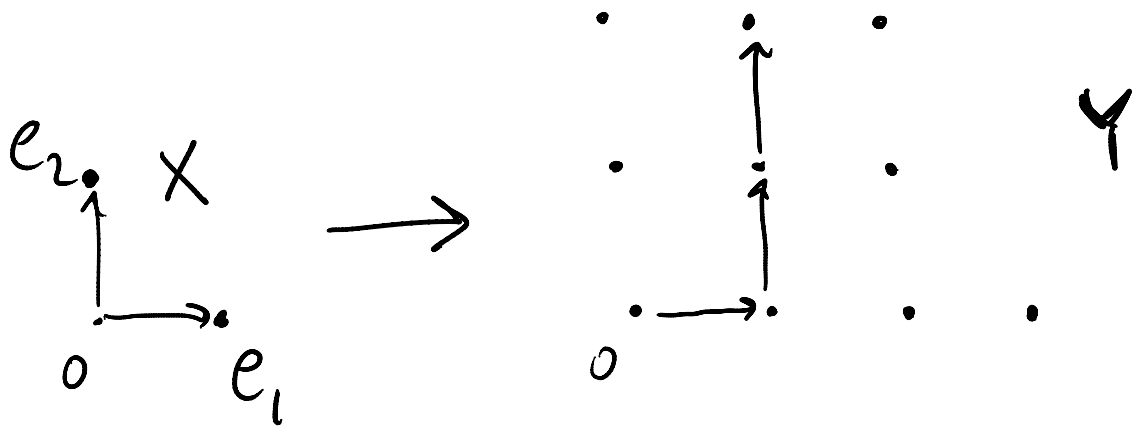


Idea: Reduction to ICA

- Basic Poissonization Lemma:

$$X_i \sim \text{Uniform}\{e_1, \ldots, e_n\}$$
$$Y = X_1 + \cdots + X_R$$
$$R \sim Poisson(\lambda)$$

Then Y has independent coordinates ($Poisson(\lambda/n)$).



Idea: Reduction to ICA

- Poissonization for Gaussian Mixture Model:

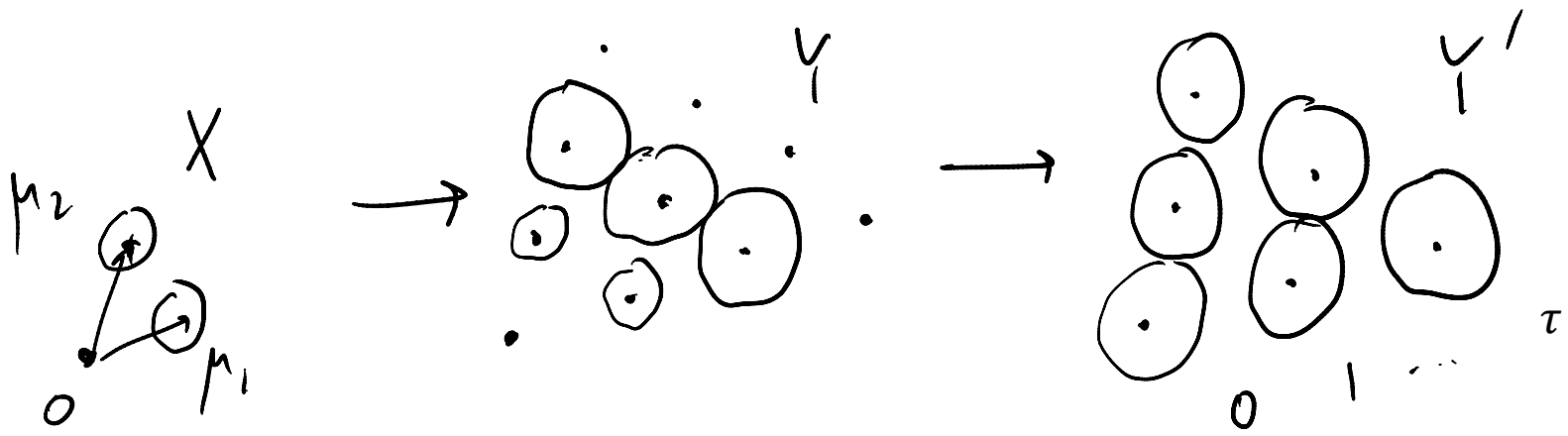
$$X \sim \mathrm{GMM} \sum_{i=1}^{k} w_i N(\mu_i, \Sigma)$$
$$Y = X_1 + \cdots + X_R$$
$$R \sim Poisson(\lambda)$$

Then Y is a linear image of a vector S having independent coordinates $Poisson(\lambda/n)$ plus noise $\eta(R)$:

$$Y = AS + \eta(R)$$

where $\eta(R) \sim \mathrm{N}(0, R^2\Sigma)$, i.e. not independent of S. $A = (\mu_1, \ldots, \mu_k)$.
Solution: pick threshold τ, reject $R \geq \tau$, add noise $\eta(\tau - R)$ to make it independent and get:

$$Y' = AS + \eta(\tau)$$



Conclusion-Summary

- Estimation of GMM hard in low dimension even for generic instances. It gets generically easier in higher dimension.