MAP Clustering under the Gaussian Mixture Model via Mixed Integer Nonlinear Programming

Patrick Flaherty

Department of Mathematics & Statistics
UMass Amherst

November 3, 2020

Outline

Table of Contents

People/Support

People

- Ji Ah Lee (UMass graduate student)
- Zhou Tang (UMass graduate student)
- Andrew Trapp (WPI)

Funding

- NSF HDR TRIPODS 1934846
- NIH 1R01GM135931-01

Motivation

- Many statistical inference problems have relevant side-information and constraints.
- Standard algorithms ignore this side-information and violate constraints.
- Ignoring the constraints can lead to inferences that don't make physical or biological sense.
- Our goal is to incorporate constraints in statistical inference and in doing so gain a deeper understanding of the tradeoffs between computational cost and statistical accuracy.
- In many cases, adding constraints improves computational efficiency due to a reduced search space.

Finite Mixture Model

- Density function for a finite mixture model:

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{y}|\boldsymbol{\theta}_k)$$

  where the observed data is $\boldsymbol{y}$ and the parameter set is $\phi = \{\boldsymbol{\theta}, \boldsymbol{\pi}\}$.
- Data is n-tuple of d-dimensional random vectors $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$.
- When the component density, $p(\boldsymbol{y}|\boldsymbol{\theta}_k)$, is a Gaussian density function, $p(\boldsymbol{y}|\phi)$ is a Gaussian mixture model with parameters $\boldsymbol{\theta} = (\{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\}, \ldots, \{\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K\})$.
- Assuming independent, identically distributed (iid) samples, the Gaussian mixture model probability density function is $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \; p\left(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$.

Gaussian Mixture Model

A generative model for the Gaussian mixture density function is

$$\begin{aligned} Z_i &\overset{\text{iid}}{\sim} \text{Categorical}(\boldsymbol{\pi}) \quad \text{for } i = 1, \ldots, n, \\ Y_i | z_i, \boldsymbol{\theta} &\sim \text{Gaussian}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}), \end{aligned} \tag{1}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$.
To generate data from the Gaussian mixture model:

1. Draw $z_i \in \{1, \ldots, K\}$ from a categorical distribution with parameter $\boldsymbol{\pi}$.
2. Given z_i, draw $\boldsymbol{y}_i$ from the associated Gaussian component distribution $p(\boldsymbol{y}_i | \boldsymbol{\theta}_{z_i})$.

MAP Clustering

- The posterior distribution function for the generative Gaussian mixture model is

$$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{\pi}) p(\boldsymbol{\theta}, \boldsymbol{\pi})}{p(\boldsymbol{y})}.$$

- The MAP clustering can be obtained by solving the following optimization problem: $\max_{\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\pi}} \log p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{y})$. s.t. $z_i \in \{1, \ldots, K\}\ \forall i$, and $\boldsymbol{\pi} \in \mathcal{P}_K$.
- In the case of one-dimensional data and equivariant components the MAP optimization problem can be written

$$\begin{aligned}
\min_{\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\pi}} \quad & \eta \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (y_i - \mu_k)^2 - \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \pi_k \\
\text{s.t.} \quad & \sum_{k=1}^{K} \pi_k = 1, \\
& \sum_{k=1}^{K} z_{ik} = 1, \qquad i = 1, \ldots, n, \\
& M_k^L \leq \mu_k \leq M_k^U, \qquad k = 1, \ldots, K, \\
& \pi_k \geq 0, \qquad k = 1, \ldots, K, \\
& z_{ik} \in \{0, 1\}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K
\end{aligned} \tag{2}$$

where $\eta = \frac{1}{2\sigma^2}$ is the precision, and M_k^L and M_k^U are real numbers.

Table of Contents

MINLPs

- Mixed integer nonlinear programming problems have both continuous and discrete variables and nonlinear functions in their objectives and constraints.

$$\begin{aligned} \min_{x,y} \quad & f(x,y) \\ \text{s.t.} \quad & g_i(x,y) = 0 \qquad i = 1,\dots,n, \\ & h_j(x,y) \leq 0 \qquad j = 1,\dots,m, \\ & x \in \mathcal{X} \subseteq \mathbb{R}^w, \\ & y \in \mathcal{Y} \subseteq \mathbb{Z}^r \end{aligned} \tag{3}$$

- MINLPs are typically solved using Generalized Benders' Decomposition or Branch-and-Bound.

MAP clustering for Gaussian Mixture Model as a Biconvex MINLP

- The GMM MAP problem can be formulated as a special kind of MINLP—a Biconvex MINLP.

$$
\begin{aligned}
\min_{\boldsymbol{z},\boldsymbol{\mu},\boldsymbol{\pi}} \quad & \eta \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}(y_i-\mu_k)^2 - \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log \pi_k \\
\text{s.t.} \quad & \sum_{k=1}^{K} \pi_k = 1, \\
& \sum_{k=1}^{K} z_{ik} = 1, \qquad i = 1,\dots,n, \\
& M_k^L \le \mu_k \le M_k^U, \qquad k = 1,\dots,K, \\
& \pi_k \ge 0, \qquad k = 1,\dots,K, \\
& z_{ik} \in \{0,1\}, \quad i = 1,\dots,n, \quad k = 1,\dots,K.
\end{aligned}
\tag{4}
$$

- If we hold $\{\boldsymbol{z},\eta\}$ fixed, the objective is convex in $\{\boldsymbol{\mu},\boldsymbol{\pi}\}$ and the constraints are linear in $\boldsymbol{\mu},\boldsymbol{\pi}$.
- If we hold $\{\boldsymbol{\mu},\boldsymbol{\pi}\}$ fixed, the objective is bilinear in $\{\boldsymbol{z},\eta\}$ and the constraints are linear in $\{\boldsymbol{z},\eta\}$.
- Note that if we separate the variables in the usual way: $\{\boldsymbol{\mu},\boldsymbol{\pi},\eta\}$ and $\boldsymbol{z}$, the problem is not biconvex.
- Biconvex problems are the subject of extensive research by Floudas and there are somewhat efficient approximation algorithms for these problems e.g. $\alpha-$ branch-and-bound.

EM Algorithm

- The EM algorithm relaxes the domain such that $z_{ik} \in [0, 1]$ instead of $z_{ik} \in \{0, 1\}$.
- The decision variables of the resulting biconvex optimization problem are partitioned into two groups: $\{\boldsymbol{z}\}$ and $\{\boldsymbol{\mu}, \boldsymbol{\pi}, \eta\}$.
- The search algorithm performs coordinate ascent on these two groups.
- There are no guarantees for the global optimality of the estimate produced by the EM algorithm.
- While the global optima of a mixture of well-separated Gaussians may have a relatively large region of attraction, inferior local optima can be arbitrarily worse than the global optimum.

Variational EM

- The variational EM algorithm introduces a surrogate function $q(\boldsymbol{z}, \phi|\boldsymbol{\xi})$ for the posterior distribution $p(\boldsymbol{z}, \boldsymbol{\phi}|\boldsymbol{y})$.
  1. the surrogate is fit to the posterior by solving $\hat{\boldsymbol{\xi}} \in \arg\min_{\boldsymbol{\xi}} \ \mathrm{KL}(q(\phi, \boldsymbol{z}|\xi) \ || \ p(\phi, \boldsymbol{z}|\boldsymbol{y}))$.
  2. the surrogate is used in place of the posterior distribution in the original optimization problem $\hat{\phi}, \hat{\boldsymbol{z}} \in \arg\min_{\phi, \boldsymbol{z}} \ \log q(\theta, \boldsymbol{z}|\boldsymbol{\xi})$.
- The search algorithm performs coordinate ascent on $\{\phi, \boldsymbol{z}\}$ and $\boldsymbol{\xi}$.
- This surrogate function approach has existed in many fields; it is alternatively known as majorization-minimization and has deep connections with Franke-Wolfe gradient methods and block coordinate descent methods.

Sequential Least Squares Programming (SLSQP)

- SLSQP is a popular general-purpose constrained nonlinear optimization method that uses a quadratic surrogate function to approximate the Lagrangian.
- In SLSQP, the surrogate function is a quadratic approximation of the Lagrangian of the original problem.
- The domain of the original problem is also relaxed so that the constraint cuts it generates are approximated by linear functions.
- Like variational EM, SLSQP iterates between fitting the surrogate function and optimizing over the decision variables.
- Quadratic surrogate functions have also been investigated in the context of variational EM for nonconjugate models.

Bandi at al.

- (Bandi et al 2019) recently described a mixed-integer optimization formulation of the parameter estimation problem for the Gaussian mixture model.
- Conditional on the parameter estimates, they computed the one-sample-at-a-time MAP assignments for out-of-sample data.
- They convincingly demonstrate that a mixed-integer optimization approach can outperform the EM algorithm in terms of out-of-sample accuracy for real-world data sets.
- Their primary objective is density estimation—to find the optimal parameters of the Gaussian mixture model. Our primary objective is MAP clustering—to find an optimal maximum a posteriori assignment of data points to clusters and associated distribution parameters.

Table of Contents

Constraints to encode prior knowledge

Many scientific studies have strict prior constraints that must not be violated in a feasible solution.

Symmetry-breaking constraint the solution is invariant to permutations,

$$\pi_1 \leq \pi_2 \leq \cdots \leq \pi_K.$$

Specific Estimators a specific estimator should be used,

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n} z_{ik}, \quad \text{and} \quad \boldsymbol{\mu}_k = \frac{\sum_{i=\mathbf{1}}^{n} \boldsymbol{y}_i z_{ik}}{\sum_{i=\mathbf{1}}^{n} z_{ik}} \; \forall k.$$

Parameter Bounds parameter that a physically impossible are not allowed

$$M_k^L \leq \mu_k M_k^U.$$

Logical Constraints replicates must cluster together: $z_{ik} = z_{jk} \forall k$; or if data point j is assigned to component k then, i must not be assigned to k: $z_{jk} \leq z_{ik}$.

Covering Constraints each components must have at least two assigned data points: $\sum_{i=1}^{n} z_{ik} \geq L, \quad \text{for } k = 1, \ldots, K.$

McCormick's Reformulation

- Recall the objective function:

$$f(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}; \boldsymbol{y}, \eta) = \eta \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} (\boldsymbol{y}_i - \mu_k)^2 - \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \pi_k.$$

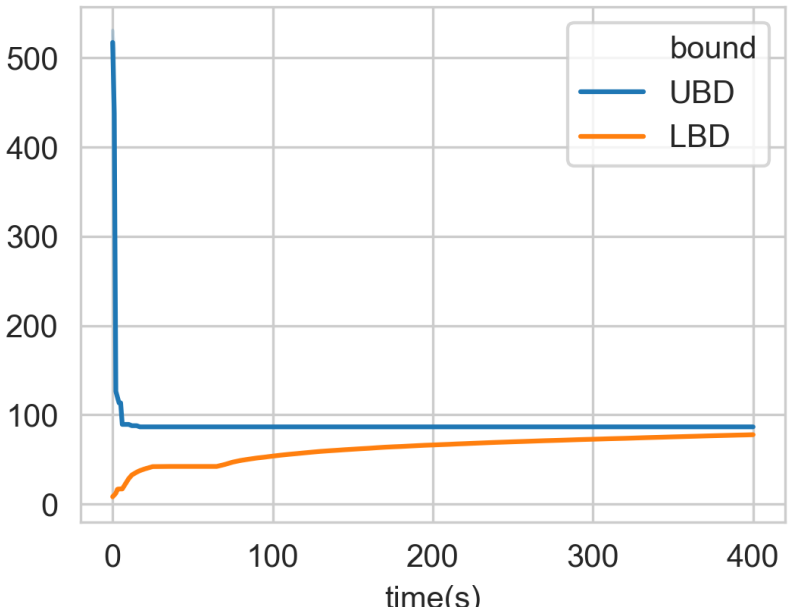
- The template matching term has two nonlinearities: $2 y_i z_{ik} \mu_k$ and $z_{ik} \mu_k^2$. These terms are frequently encountered in capital budgeting, scheduling and others.
- Given z_{ik} is a binary variable, we can rewrite the term $\sum_k z_{ik} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^2$ as $(\boldsymbol{y}_i - \sum_k z_{ik} \boldsymbol{\mu}_k)^2$ because $\sum_k z_{ik} y_i = y_i$ and each data point is constrained to be assigned to exactly one component.
- Then, we introduce a new continuous variable $t_{ik} = z_{ik} \mu_k$ which is implicitly enforced with the following four constraints for each (i, k):

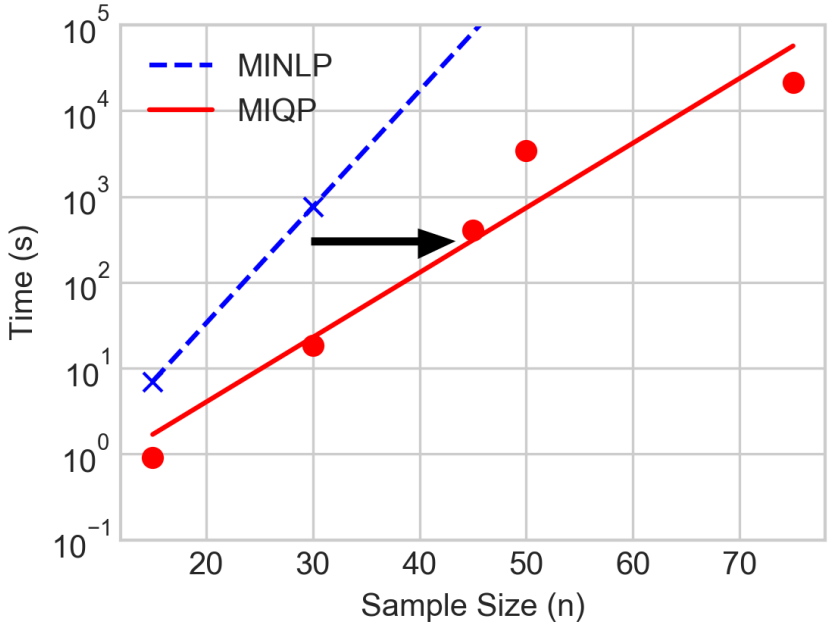
$$M_k^L z_{ik} \leq t_{ik} \leq M_k^U z_{ik},$$
$$\mu_k - M_k^U (1 - z_{ik}) \leq t_{ik} \leq \mu_k - M_k^L (1 - z_{ik}).$$

Piecewise Linear Relaxation

- The cross-entropy term, $z_{ik} \log \pi_k$, is the second source of nonlinearity.
- Approximating this nonlinearity with a piecewise linear function has two benefits:
  - the accuracy of the approximation can be controlled by the number of breakpoints in the approximation
  - sophisticated methods from ordinary and partial differential equation integration or spline fitting can be brought to service in selecting the locations of the breakpoints of the piecewise-linear approximation.
- It may be possible to set breakpoint locations adaptively as the optimization iterations progress to gain higher accuracy in the region of the MAP and the approximation can be left coarser elsewhere.

Global Convergence / Computational Efficiency





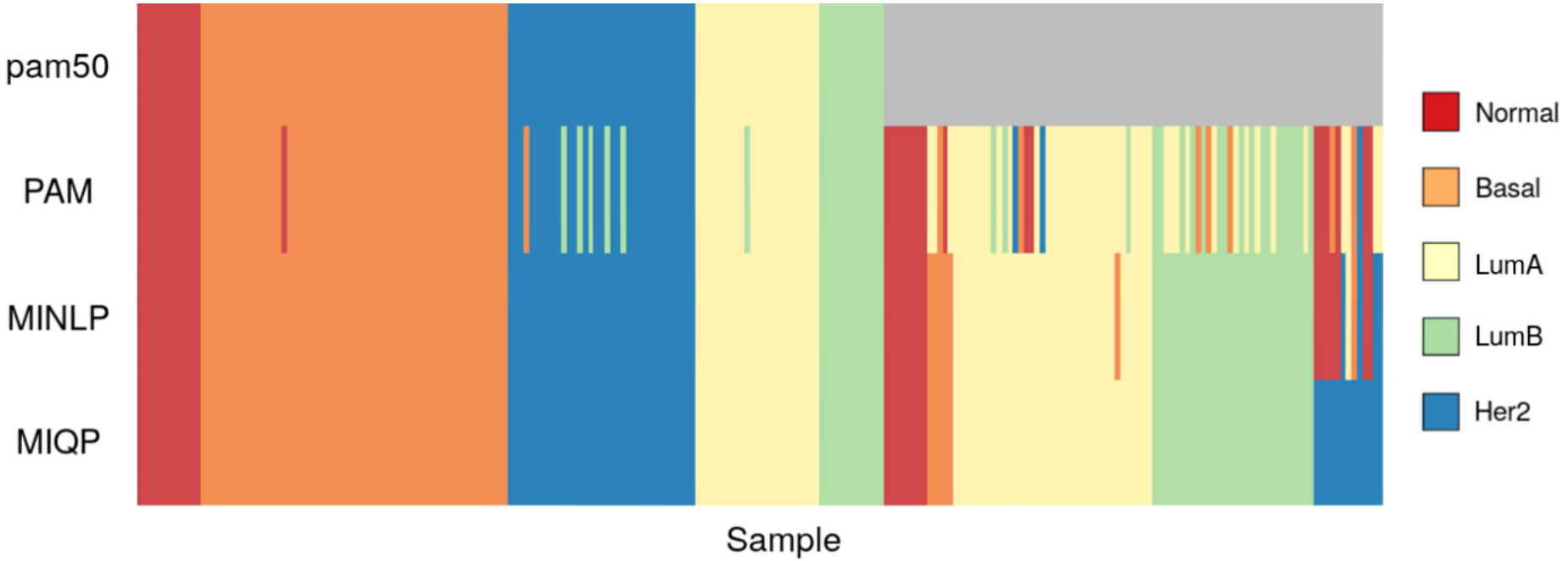
Evaluation on Standard Data Sets

		Local			Global (BnB)	
Data Set	Metric	EM	SLSQP	SA	MINLP	MIQP
iris (1 dim)	$-\log$ MAP	280.60	287.44	283.28	280.02	282.71
	LBD	—	—	—	9.27	161.60
	$\sup \lvert\hat{\pi} - \pi\rvert$	0.075	0.013	0.000	0.093	0.165
	$\lVert\hat{\mu} - \mu\rVert_2$	0.278	0.065	0.277	0.356	0.356
	$1n \sum_i \sup \lvert\hat{z}_i - z_i\rvert$	0.067	0.067	0.087	0.093	0.093
wine (13 dim)	$-\log$ MAP	1367.00	1368.71	1368.71	1366.85	1390.13
	LBD	—	—	—	-2.2e5	183.42
	$\sup \lvert\hat{\pi} - \pi\rvert$	0.005	0.066	0.066	0.006	0.167
	$\lVert\hat{\mu} - \mu\rVert_2$	2.348	1.602	1.652	1.618	14.071
	$1n \sum_i \sup \lvert\hat{z}_i - z_i\rvert$	0.006	0.006	0.006	0.006	0.022
brca (3 dim)	$-\log$ MAP	1566.49	1662.97	1662.97	1566.40	1578.49
	LBD	—	—	—	-2.7e4	332.30
	$\sup \lvert\hat{\pi} - \pi\rvert$	0.167	0.127	0.127	0.169	0.122
	$\lVert\hat{\mu} - \mu\rVert_2$	394.07	321.11	320.60	401.47	418.05
	$1n \sum_i \sup \lvert\hat{z}_i - z_i\rvert$	0.169	0.139	0.139	0.169	0.174

BRCA Expression Problem

- We evaluated our proposed approach on Prediction Analysis of Microarray 50 (pam50) gene expression data set.
- The PAM50 gene set is commonly used to identify the "intrinsic" subtypes of breast cancer among luminal A (LumA), luminal B (LumB), HER2-enriched (Her2), basal-like (Basal), and normal-like (Normal).
- Different subtypes lead to different treatment decisions, so it is critical to identify the correct subtype.
- We used the pam50 data set ($n = 232$, $d = 50$) obtained from UNC MicroArray Database.
- pam50 contains 139 subjects whose intrinsic subtypes are known, and 93 subjects whose intrinsic subtypes are unknown.

BRCA Results



Comparison of cluster assignments of our methods (MINLP, MIQP) with the PAM algorithm. For 139 samples with known instrinsic subtypes, assignments from MINLP and MIQP methods have 100% accuracy, while PAM accuracy is 94%. For the 93 samples with unknown subtypes, MINLP assignments have 68% concordance with the PAM algorithm, and MINLP has 89% concordance with MIQP assignments.

Table of Contents

Summary

- The GMM MAP clustering problem can be viewed as a biconvex mixed-integer nonlinear programming problem.
- Reformulations of the MINLP gives a MIQP optimization problem with significant computational gains.
- We can deliver better solutions for biological data sets than unconstrained clustering.

Table of Contents

PIs

Andrew McGregor (CS)

Patrick Flaherty (Stat)

Markos Katsoulakis (Math)

Arya Mazumdar (CS + EE)

Barna Saha (CS)

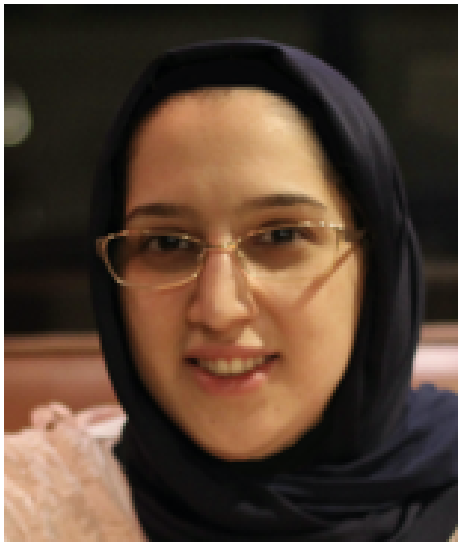
Research Areas

The overall objective is to improve theoretical understanding provide practical methods for the trade-off between computational and statistical aspects of data science problems.

1. Trade-offs between rounds of data collection and computational efficiency.
2. Minimize query complexity in interactive unsupervised learning problems.
3. Space/time complexity tradeoffs when processing stochastic data.
4. Fine-grained approximation algorithms
5. Communication-efficient distributed machine learning methods.
6. Variational inference methods with statistical guarantees given bounded computational time.
7. Principled approaches to exploit tradeoffs between bias, model complexity and computational budget.

Connect with practical problems in life sciences and physical sciences.

Postdocs



Maryam Aliakbarpour

Jeremiah Birrell

Venkata Gandikota

Tingting Zhao

REU Program

- First (to our knowledge) NSF-funded REU program in Math/Stat at UMass Amherst.
- Nathan Grant - Math/CS double major
- Joseph Cormier - US Army Reserve transfer student from local community college.

Summer Foundations of Data Science for High School Students

- In Fall 2018, Ben Marlin and I created stat/cs109f — Foundations of Data Science based on data8 at UC Berkeley.
- Planned to offer 3 week in-person course in Summer 2020 based on modules from data8, then COVID-19.
- Transitioned course to fully online with 15 students + 7 on wait list ~50% female.
- Next year, have funding for scholarships for underrepresented students in STEM.

Virtual Speaker Series

Feb 20, 2020 John Kleinberg, Cornell University

March 27, 2020 Sujay Sanghavi, UT Austin

April 17, 2020 Shachar Lovett, UCSD

May 15, 2020 Amin Karbasi, Yale

September 11, 2020 Bin Yu, UC Berkeley

November 9, 2020 Tal Rabin, UPenn

Other Activities

- Technical Workshops connecting to scientists in life sciences and physical sciences. (Spring 2021, Spring 2022)
- Theoretical computer science (TCS) Women even (Summer 2021)