Informational and Computational Limits of Clustering

and other questions about clustering

Nati Srebro
University of Toronto

based on work in progress with

Gregory Shakhnarovich
Brown University

Sam Roweis
University of Toronto

"Clustering"

- Clustering with respect to a specific model / structure / objective
- Gaussian mixture model
  - Each point comes from one of *k* "centers"
  - Gaussian cloud around each center
  - For now: unit-variance Gaussians, uniform prior over choice of center
- As an optimization problem:
  - Likelihood of centers:

$$\Sigma_i \log(\Sigma_j \exp -(x_i-\mu_j)^2/2)$$

  - *k*-means objective—Likelihood of assignment:

$$\Sigma_i \min_j (x_i-\mu_j)^2$$

Is Clustering Hard or Easy?

- *k*-means (and ML estimation?) is NP-hard
  - For some point configurations, it is hard to find the optimal solution.
  - But do these point configurations actually correspond to clusters of points?

Is Clustering Hard or Easy?

- *k*-means (and ML estimation?) is NP-hard
  - For some point configurations, it is hard to find the optimal solution.
  - But do these point configurations actually correspond to clusters of points?
- Well separated Gaussian clusters, lots of data
  - Poly time algorithms for very large separation, #points
  - Empirically, EM* works (modest separation, #points)

*EM with some bells and whistles: spectral projection (PCA), pruning centers, etc

Is Clustering Hard or Easy? (when its interesting)

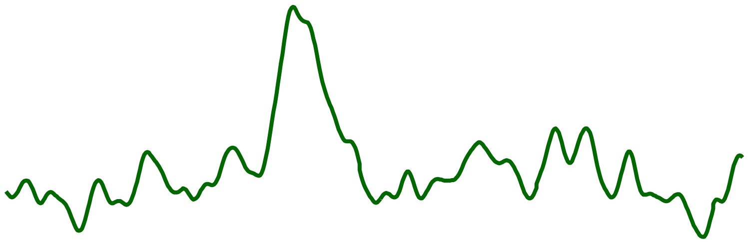
- *k*-means (and ML estimation?) is NP-hard
  - For some point configurations, it is hard to find the optimal solution.
  - But do these point configurations actually correspond to clusters of points?
- Well separated Gaussian clusters, lots of data
  - Poly time algorithms for very large separation, #points
  - Empirically, EM* works (modest separation, #points)
- Not enough data
  - Can’t identify clusters (ML clustering meaningless)

*EM with some bells and whistles: spectral projection (PCA), pruning centers, etc

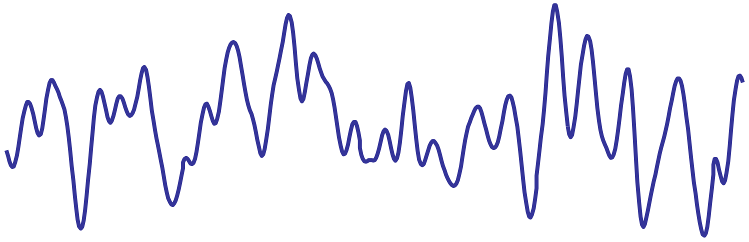
Effect of "Signal Strength"

Large separation,
More samples

Lots of data—
true solution creates
distinct peak.
Easy to find.



Not enough data—
"optimal" solution is
meaningless.

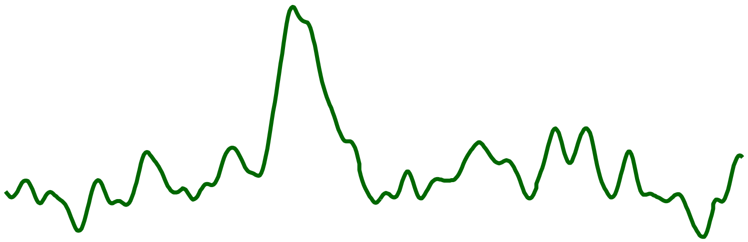


Small separation,
Less samples

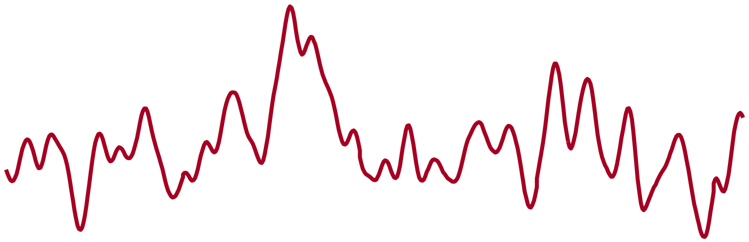
Effect of “Signal Strength”

Large separation,
More samples

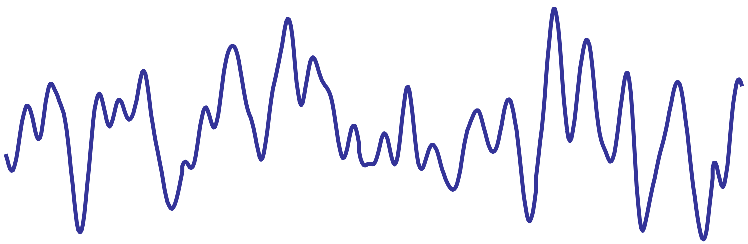
Lots of data—
true solution creates
distinct peak.
Easy to find.



Just enough data—
optimal solution is
meaningful, but hard to
find?



Not enough data—
“optimal” solution is
meaningless.



Small separation,
Less samples

Effect of “Signal Strength”

Large separation,
More samples

Lots of data—
true solution creates
distinct peak.
Easy to find.

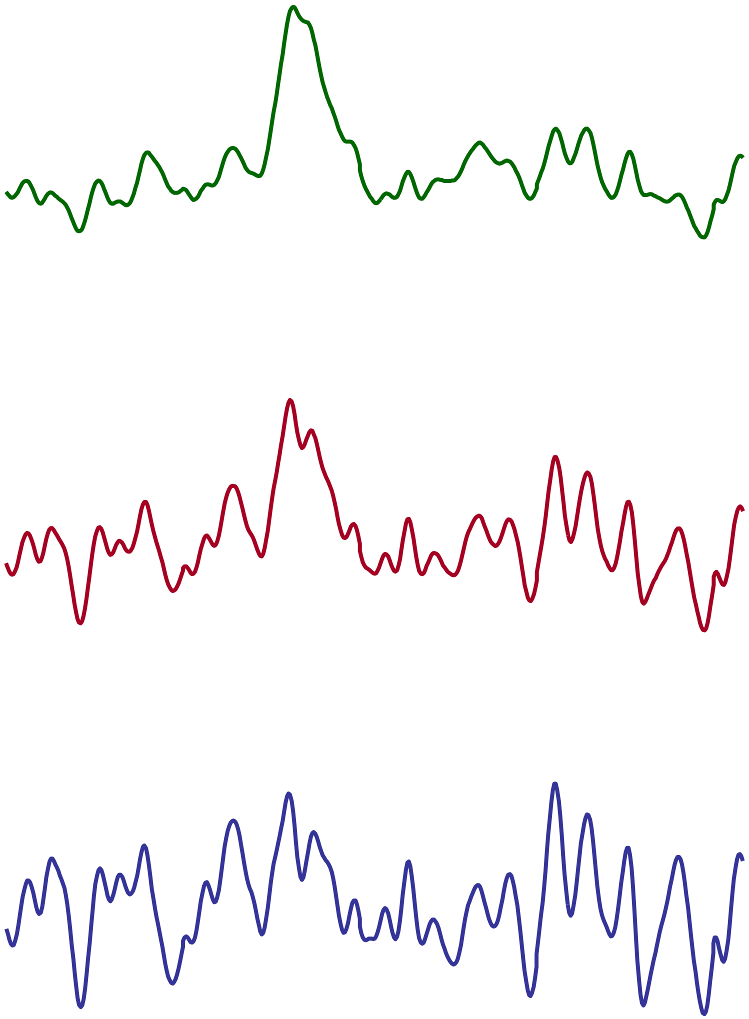
Computational
limit

Just enough data—
optimal solution is
meaningful, but hard to
find?

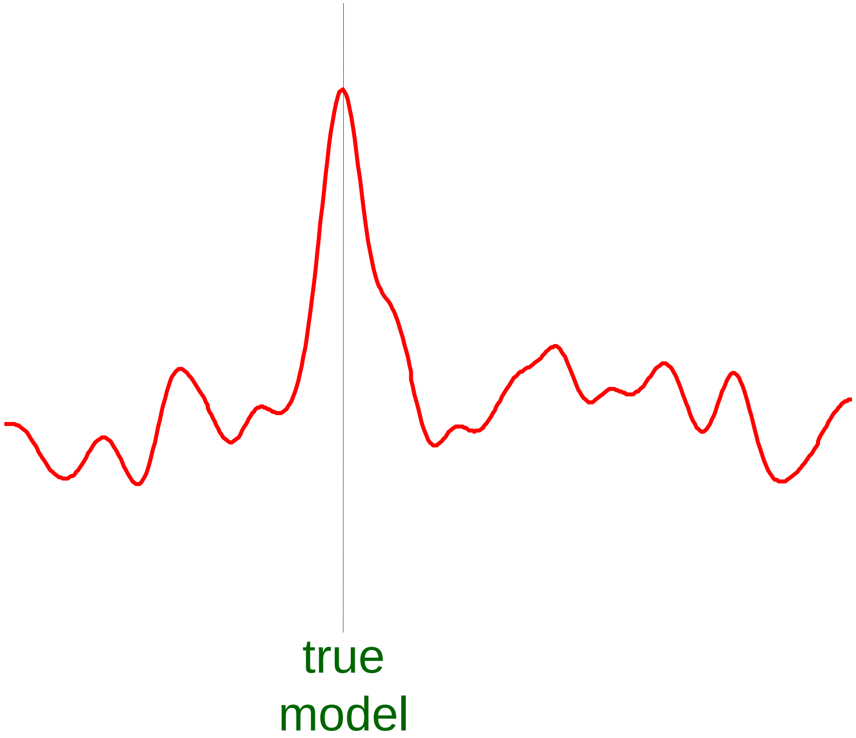
Informational
limit

Not enough data—
“optimal” solution is
meaningless.

Small separation,
Less samples



Effect of “Signal Strength”



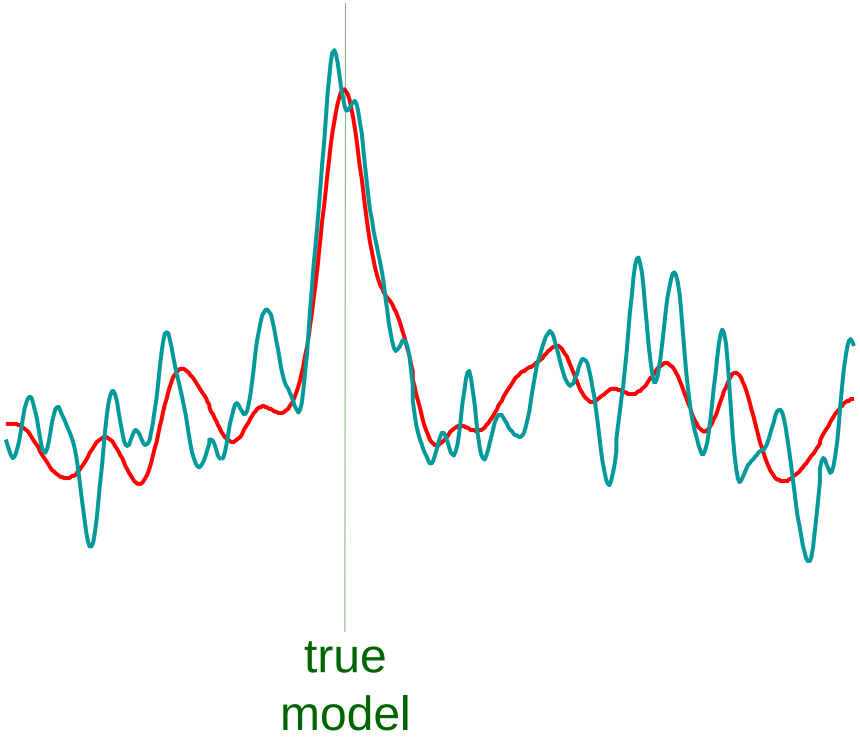
Infinite data limit:
E_x[cost(x;model)] = KL(true||model)

Mode always at true model

Determined by

- number of clusters (k)
- dimensionality (d)
- separation (s)

Effect of “Signal Strength”



Infinite data limit:
E_X[cost(x;model)] = KL(true||model)

Mode always at true model

Determined by
- number of clusters (k)
- dimensionality (d)
- separation (s)

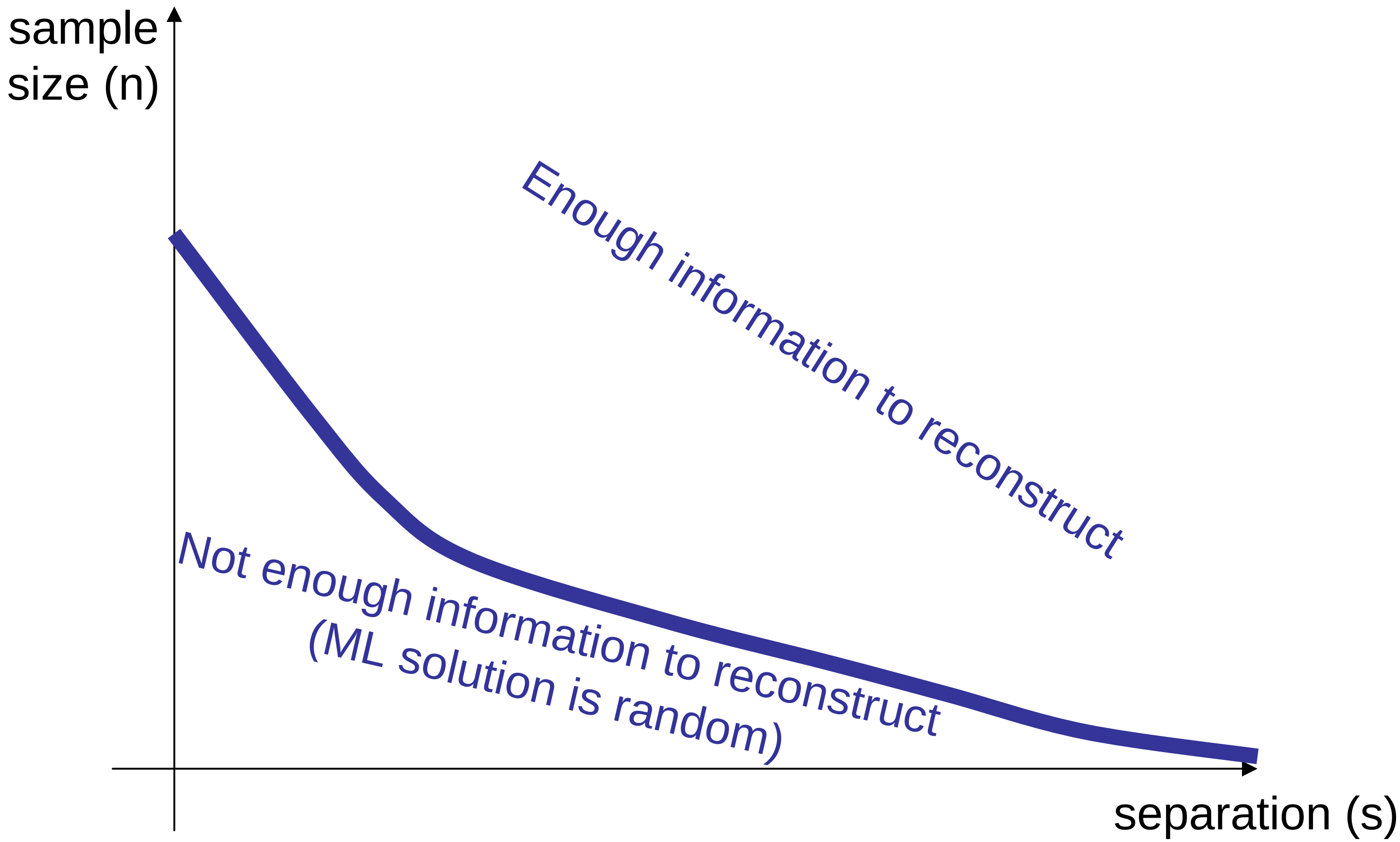
Actual log-likelihood

Also depends on:
- sample size (n)

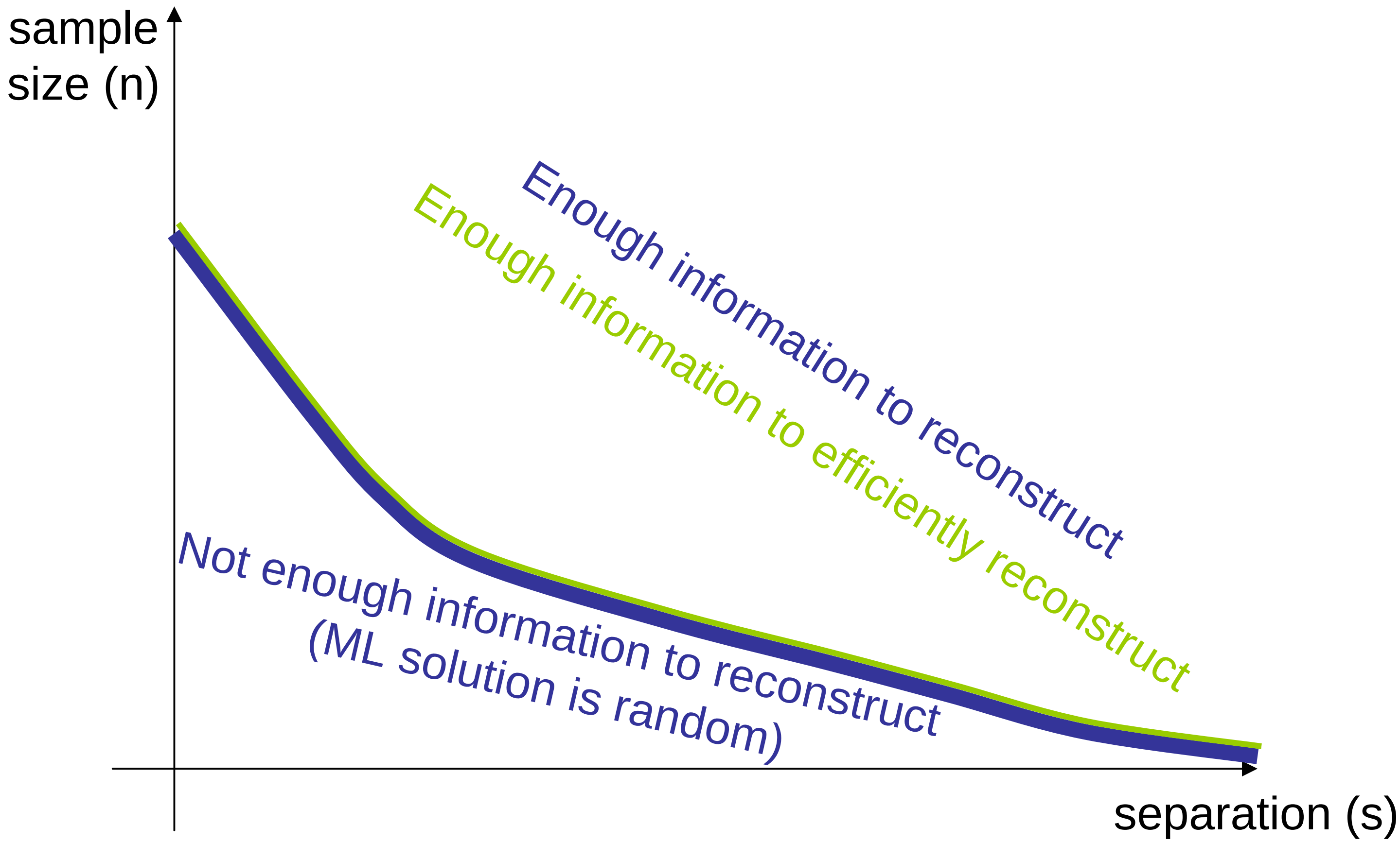
“local ML model” ~ $N(\text{true}; \frac{1}{n} J_{Fisher}^{-1})$

[Redner Walker 84]

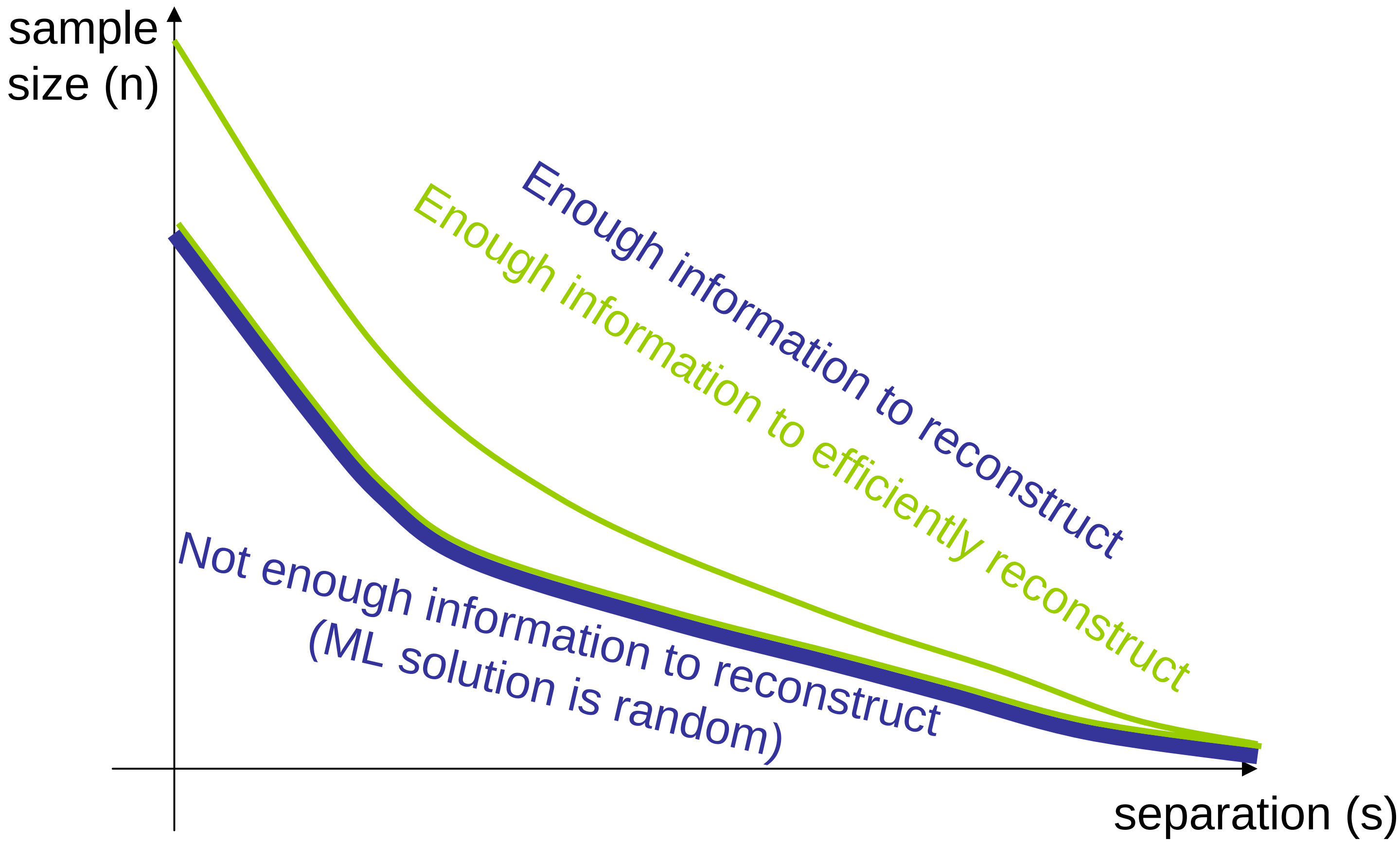
Informational and Computational Limits



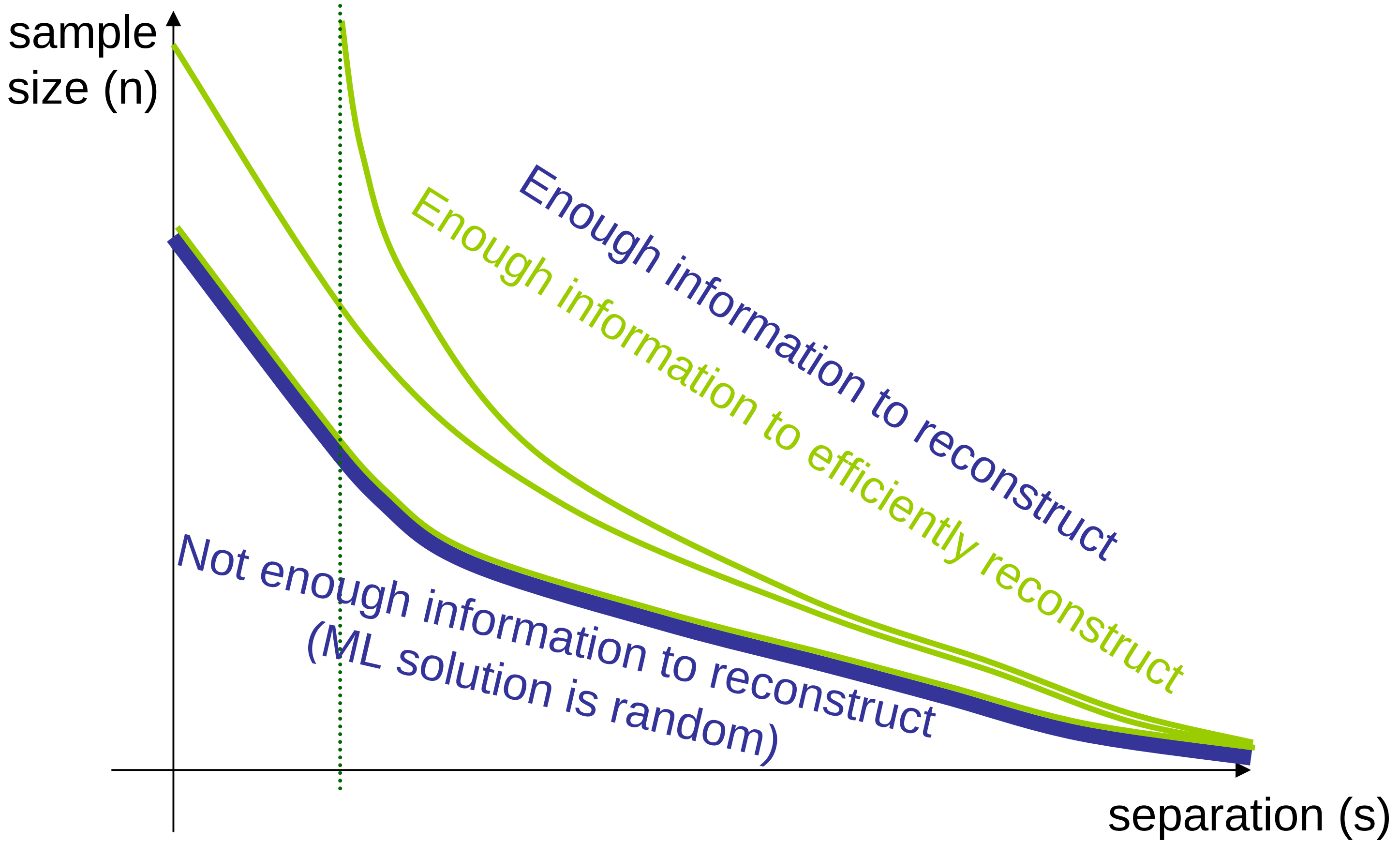
Informational and Computational Limits



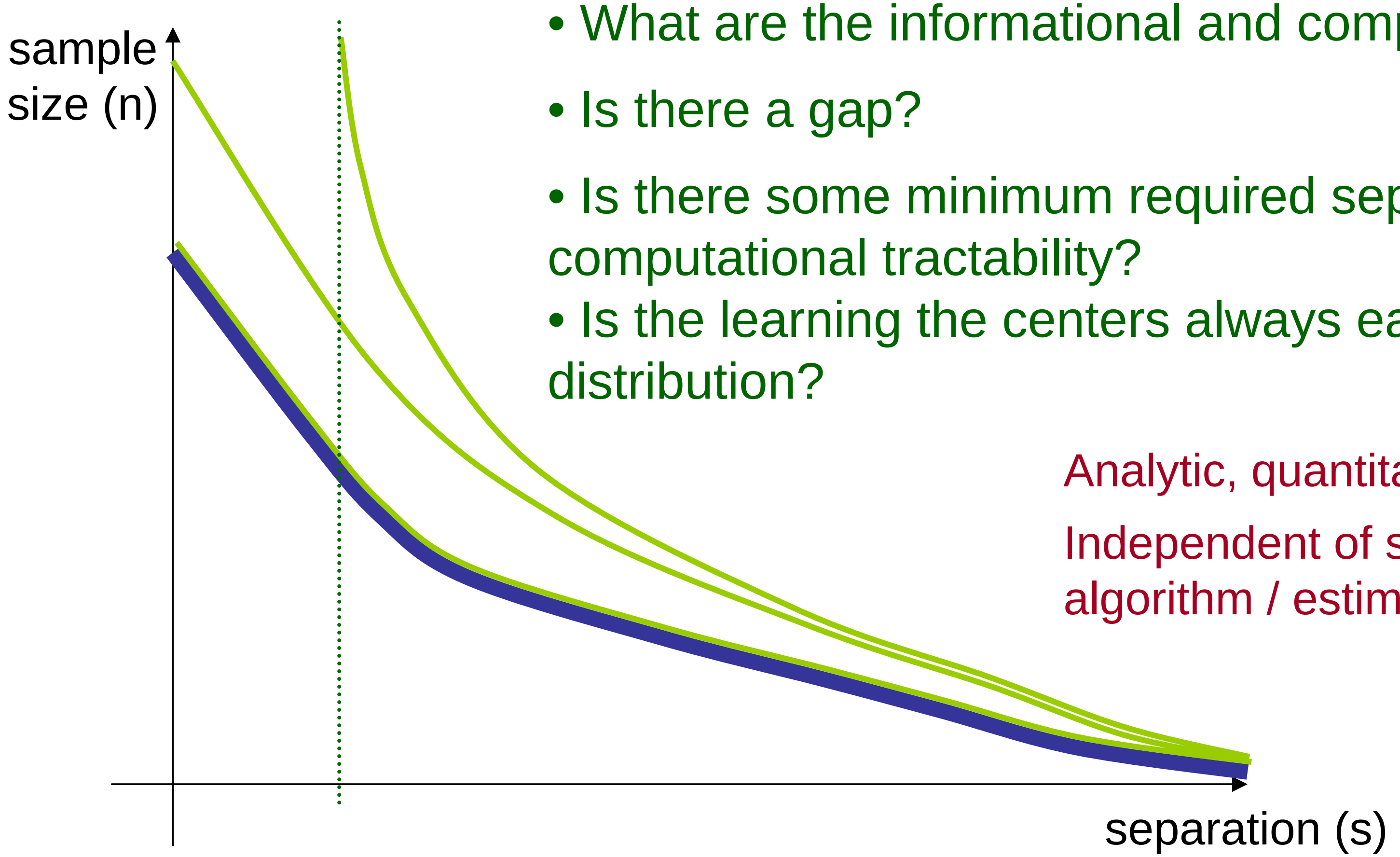
Informational and Computational Limits



Informational and Computational Limits



Informational and Computational Limits

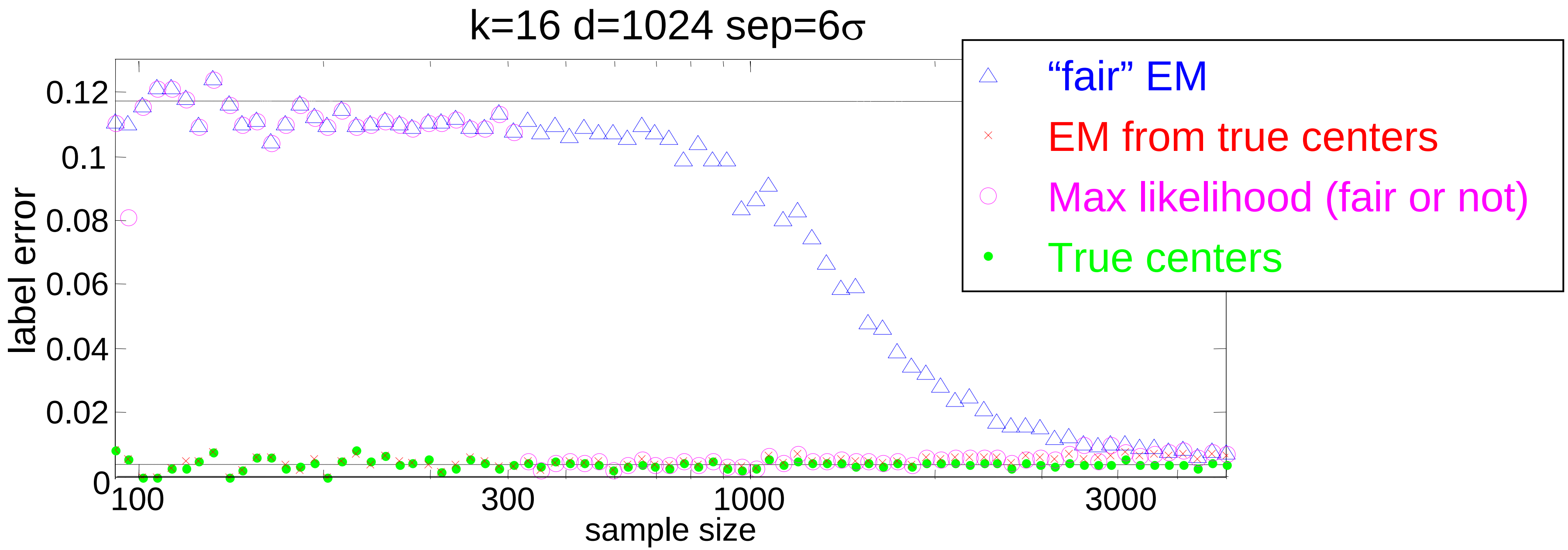


- What are the informational and computational limits?
- Is there a gap?
- Is there some minimum required separation for computational tractability?
- Is the learning the centers always easy given the true distribution?

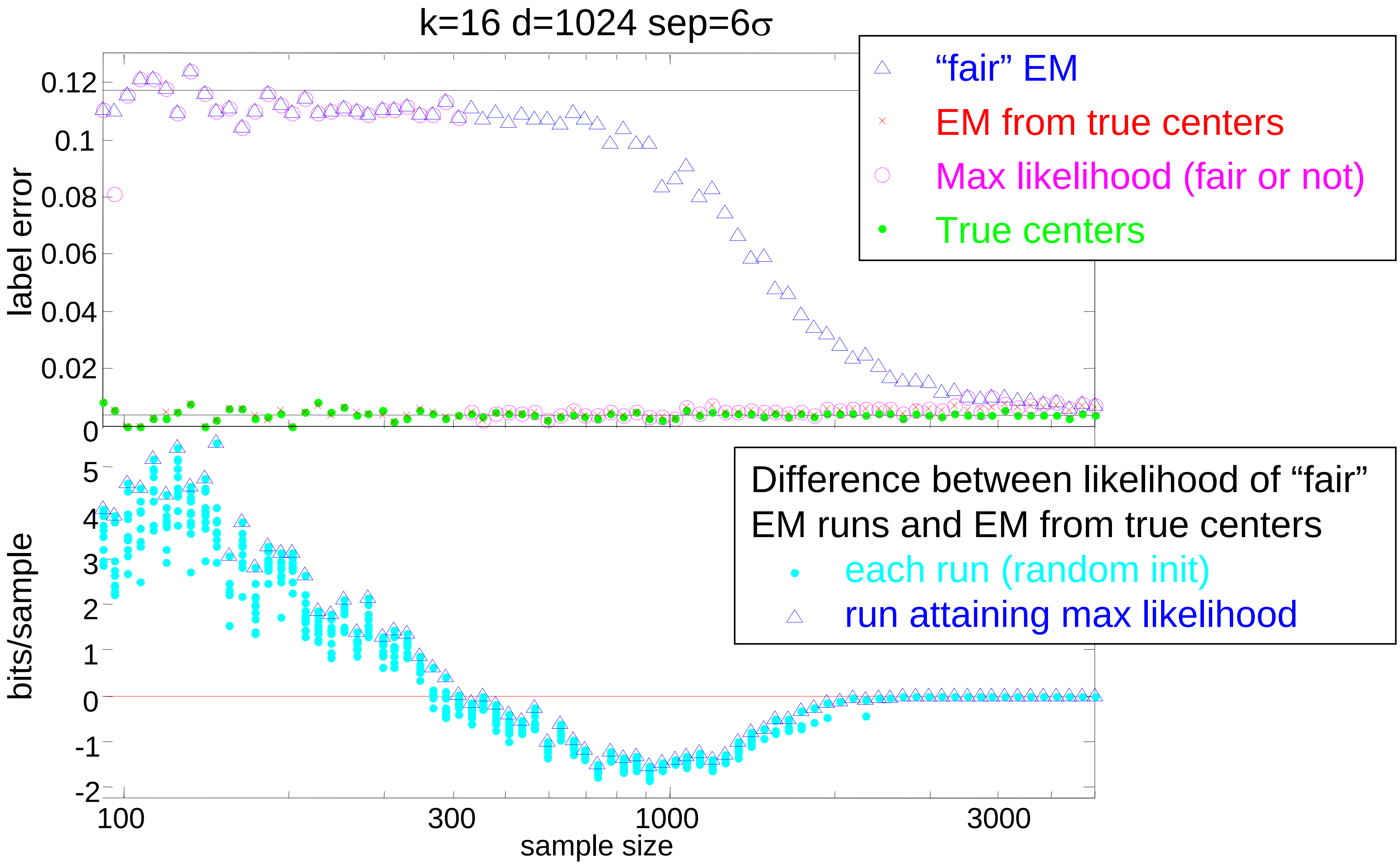
Analytic, quantitative answers.

Independent of specific algorithm / estimator

Behavior as a function of Sample Size



Behavior as a function of Sample Size



Clustering

Model of clustering

What structure are we trying to capture?
What properties do we expect the data to have?
What are we trying to get out of it?
What is a “good clustering”?

Empirical objective and evaluation

(e.g. minimization objective)

Can it be used to recover the clustering (as specified above)?
Post-hoc analysis: is what we found “real”?

Algorithm

How well does it achieve objective?
How efficient is it?
Under what circumstances?

Clustering

Model of clustering

What structure are we trying to capture?
What properties do we expect the data to have?
What are we trying to get out of it?
What is a "good clustering"?

Questions
about the world

Mathematics

Empirical objective and evaluation

(e.g. minimization objective)

Can it be used to recover the clustering (as specified above)?
Post-hoc analysis: is what we found "real"?

Algorithm

How well does it achieve objective?
How efficient is it?
Under what circumstances?

Clustering

Model of clustering

What structure are we trying to capture?
What properties do we expect the data to have?
What are we trying to get out of it?
What is a “good clustering”?

Questions about the world

Mathematics

Empirical objective and evaluation

(e.g. minimization objective)

Can it be used to recover the clustering (as specified above)?
Post-hoc analysis: is what we found “real”?

Algorithm

How well does it achieve objective?
How efficient is it?
Under what circumstances?

Clustering

Model of clustering

- What structure are we trying to capture?
- What properties do we expect the data to have?
- What are we trying to get out of it?
- What is a "good clustering"?

Questions
about the world

Mathematics

Empirical objective and evaluation

(e.g. minimization objective)

- Can it be used to recover the clustering (as specified above)?
- Post-hoc analysis: is what we found "real"?

Algorithm

- How well does it achieve objective?
- How efficient is it?
- Under what circumstances?

Clustering

Model of clustering

What structure are we trying to capture?
What properties do we expect the data to have?
What are we trying to get out of it?
What is a “good clustering”?

Questions
about the world

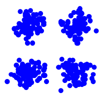
Mathematics

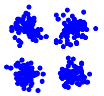
Empirical objective and evaluation

(e.g. minimization objective)

Can it be used to recover the clustering (as specified above)?
Post-hoc analysis: is what we found “real”?

Algorithm





How well does it achieve objective?
How efficient is it?
Under what circumstances?

Clustering

Model of clustering

What structure are we trying to capture?
What properties do we expect the data to have?
What are we trying to get out of it?
What is a “good clustering”?

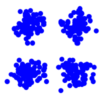
Questions
about the world

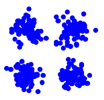
Mathematics

Empirical objective and evaluation

(e.g. minimization objective)

Can it be used to recover the clustering (as specified above)?
Post-hoc analysis: is what we found “real”?
Can what we found generalize?





Algorithm

How well does it achieve objective?
How efficient is it?
Under what circumstances?

"Clustering is Easy", take 1: Approximation Algorithms

$(1+\varepsilon)$-Approximation for k-means in time $O(2^{(k/\varepsilon)^{const}} nd)$ **[Kumar Sabharwal Sen 2004]**

For any data set of points, find clustering with k-means cost $\leq (1+\varepsilon)\times$cost-of-optimal-clustering

"Clustering is Easy", take 1: Approximation Algorithms

$(1+\varepsilon)$-Approximation for k-means in time $O(2^{(k/\varepsilon)^{const}} nd)$ [Kumar Sabharwal Sen 2004]

$\mu_1 = (\ 5,0,0,0,\ldots,0)$

$\mu_2 = (-5,0,0,0,\ldots,0)$

$0.5\ N(\mu_1,I) + 0.5\ N(\mu_2,I)$

$\text{cost}([\mu_1,\mu_2]) \approx \sum_i \min_j (x_i-\mu_j)^2 \approx d \cdot n$

$\text{cost}([0,0]) \approx \sum_i \min_j (x_i-0)^2 \approx (d+25) \cdot n$

$\Rightarrow$ [0,0] is a $(1+25/d)$-approximation

Need $\varepsilon < sep^2/d$, time becomes $O(2^{(kds)^{const}} n)$

"Clustering is Easy", take 2: Data drawn from a Gaussian Mixture

$$x_1, x_2, \ldots, x_n \sim 1/k\, N(\mu_1, \sigma^2 I) + 1/k\, N(\mu_2, \sigma^2 I) + \cdots + 1/k\, N(\mu_k, \sigma^2 I)$$

$$|\mu_i - \mu_j| > \mathbf{s} \cdot \sigma$$

- Find the modes

  (ε-neighborhood with the most points; point with closest neighbors)

  – Required sample size: $n = 2^{\Omega(d)}$

- Randomly project to $\Theta(\log k)$ dimensions
  – Now $n = \Omega(k^{\log^2 1/\delta})$ enough to find modes
  – With $s > \frac{1}{2} d^{1/2}$, modes maintained in projection

[Dasgupta 99]

“Clustering is Easy”, take 2: Data drawn from a Gaussian Mixture

$x_1, x_2, \ldots, x_n \sim 1/k\ N(\mu_1,\sigma^2 I) + 1/k\ N(\mu_2,\sigma^2 I) + \cdots + 1/k\ N(\mu_k,\sigma^2 I)$

$|\mu_i - \mu_j| > \mathbf{s} \cdot \sigma$

Dasgupta 1999	$s > 0.5 d^{1/2}$	$n = \Omega(k^{\log^2 1/\delta})$	Random projection, then mode finding
Arora Kannan 2001	$s = \Omega(d^{1/4} \log d)$		Distance based

"Clustering is Easy", take 2: Data drawn from a Gaussian Mixture

$x_1, x_2, \ldots, x_n \sim 1/k\ N(\mu_1, \sigma^2 I) + 1/k\ N(\mu_2, \sigma^2 I) + \cdots + 1/k\ N(\mu_k, \sigma^2 I)$

$|\mu_i - \mu_j| > \mathbf{s} \cdot \sigma$

- Randomly project to $\Theta(\log k)$ dimensions
  - Now $n = \Omega(k^{\log^2 1/\delta})$ enough to find modes
  - With $s > \frac{1}{2} d^{1/2}$, modes maintained in projection

[Dasgupta 99]

- Project to k principal directions (PCA)
  - Spherical Gaussian components: k principal directions of true distribution span centers
  - Required separation only $s = \Omega(k^{1/4} \log dk)$

[Vempala Wang 04]

"Clustering is Easy", take 2: Data drawn from a Gaussian Mixture

$x_1, x_2, \ldots, x_n \sim 1/k\ N(\mu_1,\sigma^2 I) + 1/k\ N(\mu_2,\sigma^2 I) + \cdots + 1/k\ N(\mu_k,\sigma^2 I)$

$|\mu_i - \mu_j| > \mathbf{s} \cdot \sigma$

Dasgupta 1999	$s > 0.5d^{1/2}$	$n = \Omega(k^{\log^2 1/\delta})$	Random projection, then mode finding	
Dagupta Schulman 2000	$s = \Omega(d^{1/4})$ (large d)	$n = \text{poly}(k)$	2 round EM with $\Theta(k \cdot \log k)$ centers	all between-class distance
Arora Kannan 2001	$s = \Omega(d^{1/4} \log d)$		Distance based	>
Vempala Wang 2004	$s = \Omega(k^{1/4} \log dk)$	$n = \Omega(d^3 k^2 \log(dk/s\delta))$	Spectral projection, then distances	all within-class distance

General mixture of Gaussians:

[Kannan Salmasian Vempala 2005] $s=\Omega(k^{5/2}\log(kd))$, $n=\Omega(k^2 d \cdot \log^5(d))$

[Achliopts McSherry 2005] $s>4k+o(k)$, $n=\Omega(k^2 d)$