# Disentangling Gaussians

Ankur Moitra, MIT

November 6th, 2014 — Dean's Breakfast

# The Gaussian Distribution

The Gaussian distribution is defined as ($\mu$ = mean, $\sigma^2$ = variance):

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# The Gaussian Distribution

The Gaussian distribution is defined as ($\mu$ = mean, $\sigma^2$ = variance):

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**Central Limit Theorem:** The sum of independent random variables $X_1, X_2, ...., X_s$ converges in distribution to a Gaussian:

$$\frac{1}{\sqrt{s}} \sum_{i=1}^{s} X_i \rightarrow_d \mathcal{N}(\mu, \sigma^2)$$

# The Gaussian Distribution

The Gaussian distribution is defined as ($\mu$ = mean, $\sigma^2$ = variance):

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

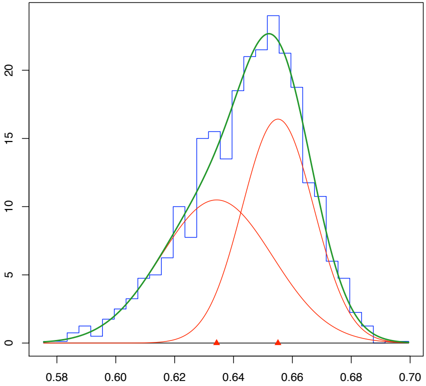
**Central Limit Theorem:** The sum of independent random variables $X_1, X_2, ...., X_s$ converges in distribution to a Gaussian:

$$\frac{1}{\sqrt{s}} \sum_{i=1}^{s} X_i \to_d \mathcal{N}(\mu, \sigma^2)$$

This distribution is ubiquitous — e.g. used to model height, velocities in an ideal gas, annual rainfall, ...

# Karl Pearson (1894) and the Naples Crabs

(figure due to Peter Macdonald)



# Gaussian Mixture Models

$$F(x) = w_1 F_1(x) + (1 - w_1) F_2(x), \text{ where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$$

In particular, with probability $w_1$ output a sample from $F_1$, otherwise output a sample from $F_2$

# Gaussian Mixture Models

$$F(x) = w_1 F_1(x) + (1 - w_1) F_2(x), \text{ where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$$

In particular, with probability $w_1$ output a sample from $F_1$, otherwise output a sample from $F_2$

five unknowns: $w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$

# Gaussian Mixture Models

$$F(x) = w_1 F_1(x) + (1 - w_1) F_2(x), \text{ where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$$

In particular, with probability $w_1$ output a sample from $F_1$, otherwise output a sample from $F_2$

five unknowns: $w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$

**Question**

*Given enough random samples from $F$, can we learn these parameters (approximately)?*

# Gaussian Mixture Models

$$F(x) = w_1 F_1(x) + (1 - w_1) F_2(x), \text{ where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$$

In particular, with probability $w_1$ output a sample from $F_1$, otherwise output a sample from $F_2$

five unknowns: $w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$

**Question**

*Given enough random samples from $F$, can we learn these parameters (approximately)?*

Pearson invented the **method of moments**, to attack this problem...

# Pearson's Sixth Moment Test

**Claim**

$E_{x \leftarrow F(x)}[x^r]$ *is a polynomial in* $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

# Pearson's Sixth Moment Test

**Claim**

*$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$*

In particular:

- $E_{x \leftarrow F(x)}[x] = w_1\mu_1 + (1 - w_1)\mu_2$

# Pearson's Sixth Moment Test

**Claim**

*$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$*

In particular:

- $E_{x \leftarrow F(x)}[x] = w_1\mu_1 + (1 - w_1)\mu_2$
- $E_{x \leftarrow F(x)}[x^2] = w_1(\mu_1^2 + \sigma_1^2) + (1 - w_1)(\mu_2^2 + \sigma_2^2)$

# Pearson's Sixth Moment Test

**Claim**

$E_{x \leftarrow F(x)}[x^r]$ *is a polynomial in* $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

In particular:

- $E_{x \leftarrow F(x)}[x] = w_1\mu_1 + (1 - w_1)\mu_2$
- $E_{x \leftarrow F(x)}[x^2] = w_1(\mu_1^2 + \sigma_1^2) + (1 - w_1)(\mu_2^2 + \sigma_2^2)$
- $E_{x \leftarrow F(x)}[x^3] = w_1(\mu_1^3 + 3\mu_1\sigma_1^2) + (1 - w_1)(\mu_2^3 + 3\mu_2\sigma_2^2)$

# Pearson's Sixth Moment Test

**Claim**

*$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$*

In particular:

- $E_{x \leftarrow F(x)}[x] = w_1\mu_1 + (1 - w_1)\mu_2$
- $E_{x \leftarrow F(x)}[x^2] = w_1(\mu_1^2 + \sigma_1^2) + (1 - w_1)(\mu_2^2 + \sigma_2^2)$
- $E_{x \leftarrow F(x)}[x^3] = w_1(\mu_1^3 + 3\mu_1\sigma_1^2) + (1 - w_1)(\mu_2^3 + 3\mu_2\sigma_2^2)$
- ...

# Pearson's Sixth Moment Test

**Claim**

*$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$*

Let $E_{x \leftarrow F(x)}[x^r] = M_r(\theta)$

# Pearson's Sixth Moment Test

**Claim**

*$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$*

Let $E_{x \leftarrow F(x)}[x^r] = M_r(\theta)$

# Pearson's Sixth Moment Test

**Claim**

$E_{x \leftarrow F(x)}[x^r]$ *is a polynomial in* $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

Let $E_{x \leftarrow F(x)}[x^r] = M_r(\theta)$

---

- Gather samples $S$

---

# Pearson's Sixth Moment Test

**Claim**

$E_{x \leftarrow F(x)}[x^r]$ *is a polynomial in* $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

Let $E_{x \leftarrow F(x)}[x^r] = M_r(\theta)$

---

- Gather samples $S$
- Set $\widetilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ for $r = 1, 2, ..., 6$

---

# Pearson's Sixth Moment Test

**Claim**

*$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$*

Let $E_{x \leftarrow F(x)}[x^r] = M_r(\theta)$

---

- Gather samples $S$
- Set $\widetilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ for $r = 1, 2, ..., 6$
- Compute simultaneous roots of $\{M_r(\theta) = \widetilde{M}_r\}_{r=1,2,...,5}$,

---

# Pearson's Sixth Moment Test

**Claim**

*$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$*

Let $E_{x \leftarrow F(x)}[x^r] = M_r(\theta)$

---

- Gather samples $S$
- Set $\widetilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ for $r = 1, 2, ..., 6$
- Compute simultaneous roots of $\{M_r(\theta) = \widetilde{M}_r\}_{r=1,2,...,5}$, select root $\theta$ that is closest in **sixth** moment

---

# Provable Guarantees?

In Contributions to the Mathematical Theory of Evolution (attributed to George Darwin):

> *"Given the probable error of every ordinate of a frequency curve, what are the probable errors of the elements of the two normal curves into which it may be dissected?"*

# Provable Guarantees?

In <u>Contributions to the Mathematical Theory of Evolution</u> (attributed to George Darwin):

> *"Given the probable error of every ordinate of a frequency curve, what are the probable errors of the elements of the two normal curves into which it may be dissected?"*

---

- Are the parameters of a mixture of two Gaussians uniquely determined by its moments?

# Provable Guarantees?

In Contributions to the Mathematical Theory of Evolution (attributed to George Darwin):

> *"Given the probable error of every ordinate of a frequency curve, what are the probable errors of the elements of the two normal curves into which it may be dissected?"*

---

- Are the parameters of a mixture of two Gaussians uniquely determined by its moments?
- Are these polynomial equations robust to errors?

# A View from Theoretical Computer Science

Suppose our goal is to **provably** learn the parameters of each component within an additive $\epsilon$:

Goal

*Output a mixture* $\widehat{F} = \widehat{w}_1 \widehat{F}_1 + \widehat{w}_2 \widehat{F}_2$ *so that there is a permutation* $\pi : \{1, 2\} \rightarrow \{1, 2\}$ *and for* $i = \{1, 2\}$

$$|w_i - \widehat{w}_{\pi(i)}|, |\mu_i - \widehat{\mu}_{\pi(i)}|, |\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon$$

# A View from Theoretical Computer Science

Suppose our goal is to **provably** learn the parameters of each component within an additive $\epsilon$:

### Goal

*Output a mixture* $\widehat{F} = \widehat{w}_1 \widehat{F}_1 + \widehat{w}_2 \widehat{F}_2$ *so that there is a permutation* $\pi : \{1,2\} \to \{1,2\}$ *and for* $i = \{1,2\}$

$$|w_i - \widehat{w}_{\pi(i)}|, |\mu_i - \widehat{\mu}_{\pi(i)}|, |\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon$$

Is there an algorithm that takes poly$(1/\epsilon)$ samples and runs in time poly$(1/\epsilon)$?

# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)

# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)

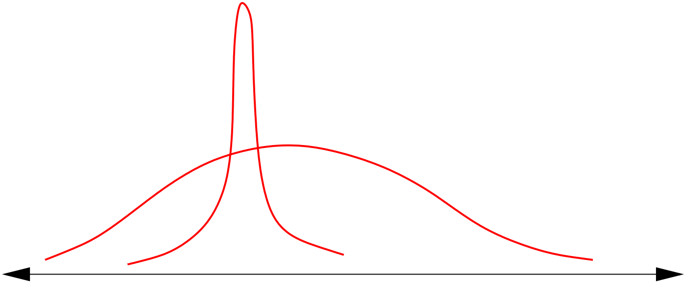
# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**

# A Conceptual History

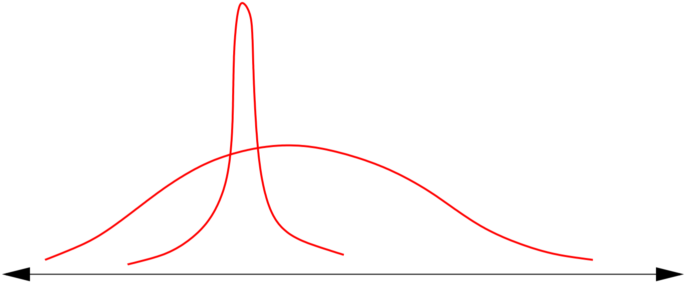
- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails

# Identifiability through the Tails



**Approach:** Find the parameters of the component with largest variance (it dominates the behavior of $F(x)$ at infinity)

# Identifiability through the Tails



**Approach:** Find the parameters of the component with largest variance (it dominates the behavior of $F(x)$ at infinity); subtract it off and continue

# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails

# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails
  requires **many** samples

# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails
  requires **many** samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM)

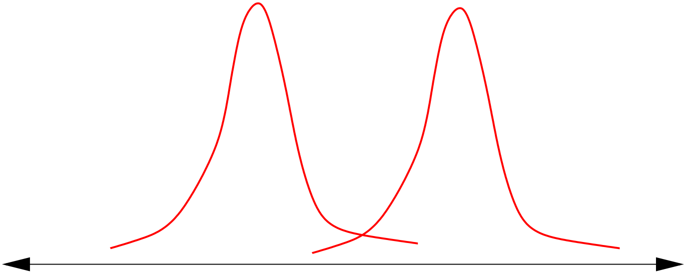
# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails
  requires **many** samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM)
  gets stuck in **local maxima**

# A Conceptual History

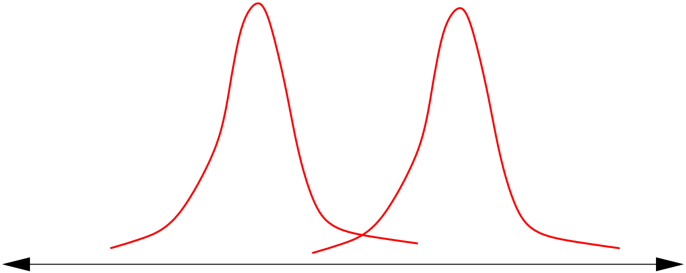
- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails
  requires **many** samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM)
  gets stuck in **local maxima**
- Dasgupta (1999) and many others: Clustering

# Clustering Well-separated Mixtures



**Approach:** Cluster samples based on which component generated them

# Clustering Well-separated Mixtures



**Approach:** Cluster samples based on which component generated them; output the empirical mean and variance of each cluster

# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails
  requires **many** samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM)
  gets stuck in **local maxima**
- Dasgupta (1999) and many others: Clustering

# A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
  consistent and efficient, usually **computationally hard**
- Teicher (1961): Identifiability through tails
  requires **many** samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM)
  gets stuck in **local maxima**
- Dasgupta (1999) and many others: Clustering
  assumes almost **non-overlapping** components

In summary, these approaches are heuristic, computationally intractable or make a separation assumption about the mixture

In summary, these approaches are heuristic, computationally intractable or make a separation assumption about the mixture

**Question**

*What if the components overlap almost entirely?*

In summary, these approaches are heuristic, computationally intractable or make a separation assumption about the mixture

Question

*What if the components overlap almost entirely?*

**[Kalai, Moitra, Valiant]** (studies $n$-dimensional version too):

- Reduce to the one-dimensional case

In summary, these approaches are heuristic, computationally intractable or make a separation assumption about the mixture

**Question**

*What if the components overlap almost entirely?*

**[Kalai, Moitra, Valiant]** (studies $n$-dimensional version too):

- Reduce to the one-dimensional case
- Analyze Pearson's sixth moment test (with noisy estimates)

# Our Results

Suppose $w_1 \in [\epsilon^{10}, 1 - \epsilon^{10}]$ and $\int |F_1(x) - F_2(x)| dx \geq \epsilon^{10}$

# Our Results

Suppose $w_1 \in [\epsilon^{10}, 1 - \epsilon^{10}]$ and $\int |F_1(x) - F_2(x)| dx \geq \epsilon^{10}$

**Theorem (Kalai, Moitra, Valiant)**

*There is an algorithm that (with probability at least $1 - \delta$) learns the parameters of $F$ within an additive $\epsilon$, and the running time and number of samples needed are poly$(\frac{1}{\epsilon}, \log \frac{1}{\delta})$.*

# Our Results

Suppose $w_1 \in [\epsilon^{10}, 1 - \epsilon^{10}]$ and $\int |F_1(x) - F_2(x)| dx \geq \epsilon^{10}$

**Theorem (Kalai, Moitra, Valiant)**

*There is an algorithm that (with probability at least $1 - \delta$) learns the parameters of $F$ within an additive $\epsilon$, and the running time and number of samples needed are poly$(\frac{1}{\epsilon}, \log \frac{1}{\delta})$.*

Previously, the best known bound on the running time/sample complexity were exponential

# Our Results

Suppose $w_1 \in [\epsilon^{10}, 1 - \epsilon^{10}]$ and $\int |F_1(x) - F_2(x)| dx \geq \epsilon^{10}$

**Theorem (Kalai, Moitra, Valiant)**

*There is an algorithm that (with probability at least $1 - \delta$) learns the parameters of $F$ within an additive $\epsilon$, and the running time and number of samples needed are poly$(\frac{1}{\epsilon}, \log \frac{1}{\delta})$.*

Previously, the best known bound on the running time/sample complexity were exponential

See also **[Moitra, Valiant]** and **[Belkin, Sinha]** for mixtures of $k$ Gaussians

# Analyzing the Method of Moments

Let's start with an easier question:

# Analyzing the Method of Moments

Let's start with an easier question:

**Question**

*What if we are given the first* **six** *moments of the mixture, exactly?*

# Analyzing the Method of Moments

Let's start with an easier question:

**Question**
*What if we are given the first* **six** *moments of the mixture, exactly?*

Does this uniquely determine the parameters of the mixture?

(up to a relabeling of the components)

# Analyzing the Method of Moments

Let's start with an easier question:

**Question**

*What if we are given the first* **six** *moments of the mixture, exactly?*

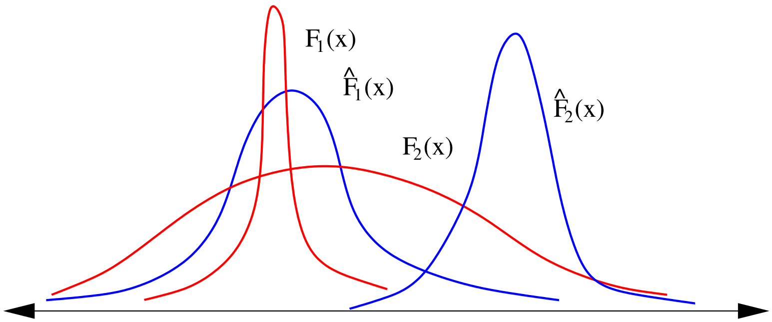
Does this uniquely determine the parameters of the mixture?

(up to a relabeling of the components)

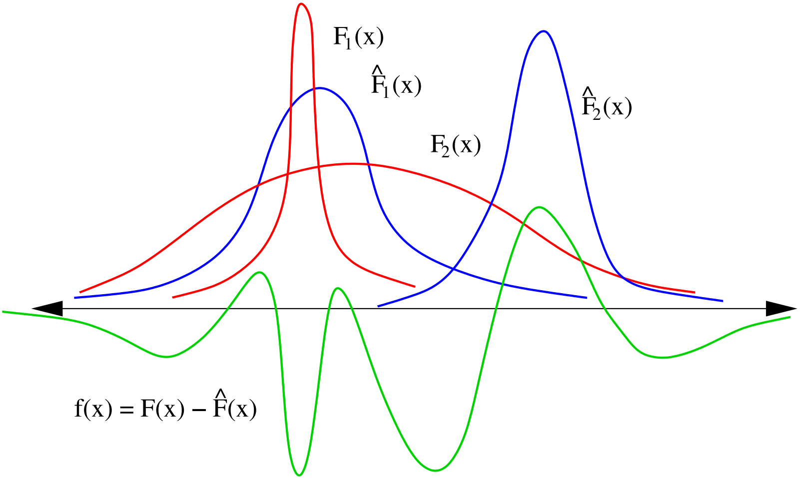
**Question**

*Do any two different mixtures $F$ and $\widehat{F}$ differ on at least one of the first six moments?*

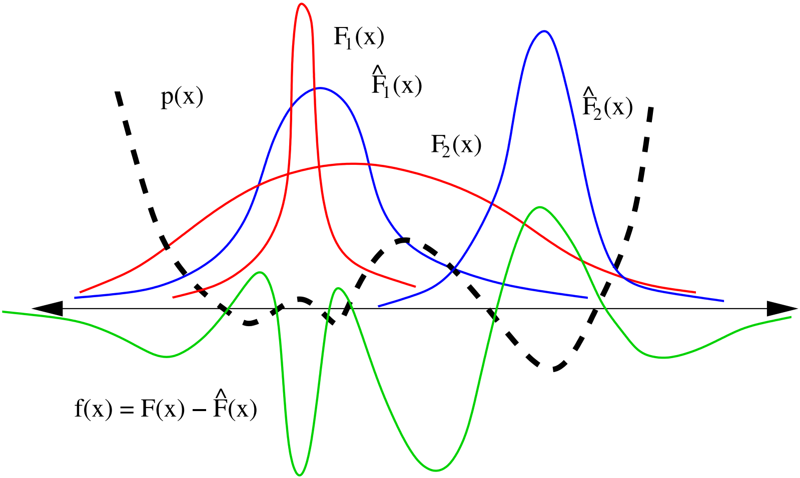
# Method of Moments



# Method of Moments



# Method of Moments



### Claim

*One of the first six moment of $F, \widehat{F}$ is different!*

## Claim

*One of the first six moment of $F, \widehat{F}$ is different!*

**Proof:**

$$0 < \left| \int_x p(x) f(x) dx \right| \;=\; \left| \int_x \sum_{r=1}^{6} p_r x^r f(x) dx \right|$$

### Claim

*One of the first six moment of $F, \widehat{F}$ is different!*

**Proof:**

$$\begin{aligned} 0 < \left| \int_x p(x) f(x) dx \right| &= \left| \int_x \sum_{r=1}^{6} p_r x^r f(x) dx \right| \\ &\leq \sum_{r=1}^{6} |p_r| \left| \int_x x^r f(x) dx \right| \end{aligned}$$

## Claim

*One of the first six moment of $F, \widehat{F}$ is different!*

**Proof:**

$$
\begin{aligned}
0 < \Big|\int_x p(x)f(x)dx\Big| &= \Big|\int_x \sum_{r=1}^{6} p_r x^r f(x)dx\Big| \\
&\leq \sum_{r=1}^{6} |p_r| \Big|\int_x x^r f(x)dx\Big| \\
&= \sum_{r=1}^{6} |p_r| |M_r(F) - M_r(\widehat{F})|
\end{aligned}
$$

## Claim

*One of the first six moment of $F, \widehat{F}$ is different!*

**Proof:**

$$\begin{aligned} 0 < \left| \int_x p(x) f(x) dx \right| &= \left| \int_x \sum_{r=1}^{6} p_r x^r f(x) dx \right| \\ &\leq \sum_{r=1}^{6} |p_r| \left| \int_x x^r f(x) dx \right| \\ &= \sum_{r=1}^{6} |p_r| |M_r(F) - M_r(\widehat{F})| \end{aligned}$$

So $\exists_{r \in \{1,2,\ldots,6\}}$ such that $|M_r(F) - M_r(\widehat{F})| > 0$

Our goal is to prove the following:

### Proposition

*If* $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ *is not identically zero,* $f(x)$ *has at most* $2k - 2$ *zero crossings (*$\alpha_i$*'s can be negative).*

Our goal is to prove the following:

**Proposition**

*If* $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ *is not identically zero,* $f(x)$ *has at most* $2k - 2$ *zero crossings (*$\alpha_i$*'s can be negative).*

......

We will do it through properties of the **heat equation**

......

# The Heat Equation

**Question**

*If the initial heat distribution on a one-dimensional infinite rod* $(\kappa)$ *is* $f(x) = f(x, 0)$ *what is the heat distribution* $f(x, t)$ *at time* $t$*?*

# The Heat Equation

**Question**

*If the initial heat distribution on a one-dimensional infinite rod ($\kappa$) is $f(x) = f(x, 0)$ what is the heat distribution $f(x, t)$ at time $t$?*

There is a probabilistic interpretation ($\sigma^2 = 2\kappa t$):

$$f(x, t) = \mathbb{E}_{z \leftarrow \mathcal{N}(0, \sigma^2)}[f(x + z, 0)]$$

# The Heat Equation

**Question**

*If the initial heat distribution on a one-dimensional infinite rod ($\kappa$) is $f(x) = f(x, 0)$ what is the heat distribution $f(x, t)$ at time $t$?*

There is a probabilistic interpretation ($\sigma^2 = 2\kappa t$):

$$f(x, t) = \mathbb{E}_{z \leftarrow \mathcal{N}(0,\sigma^2)}[f(x + z, 0)]$$

Alternatively, this is called a **convolution**:

$$f(x, t) = \int_{z=-\infty}^{\infty} f(x + z)\mathcal{N}(0, \sigma^2, z)dz = f(x) * \mathcal{N}(0, \sigma^2)$$

# The Key Facts

**Theorem (Hummel, Gidas)**

*Suppose $f(x) : \mathbb{R} \to \mathbb{R}$ is analytic and has $N$ zeros. Then*

$$f(x) * \mathcal{N}(0, \sigma^2, x)$$

*has at most $N$ zeros (for any $\sigma^2 > 0$).*

# The Key Facts

**Theorem (Hummel, Gidas)**

*Suppose $f(x) : \mathbb{R} \to \mathbb{R}$ is analytic and has $N$ zeros. Then*

$$f(x) * \mathcal{N}(0, \sigma^2, x)$$

*has at most $N$ zeros (for any $\sigma^2 > 0$).*

Convolving by a Gaussian does not increase # of zero crossings

# The Key Facts

**Theorem (Hummel, Gidas)**

*Suppose* $f(x) : \mathbb{R} \to \mathbb{R}$ *is analytic and has* $N$ *zeros. Then*

$$f(x) * \mathcal{N}(0, \sigma^2, x)$$

*has at most* $N$ *zeros (for any* $\sigma^2 > 0$*).*

Convolving by a Gaussian does not increase # of zero crossings

**Fact**

$$\mathcal{N}(\mu_1, \sigma_1^2, x) * \mathcal{N}(\mu_2, \sigma_2^2, x) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2, x)$$

Recall, our goal is to prove the following:

### Proposition

*If $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ is not identically zero, $f(x)$ has at most $2k-2$ zero crossings ($\alpha_i$'s can be negative).*

Recall, our goal is to prove the following:

### Proposition

*If $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ is not identically zero, $f(x)$ has at most $2k - 2$ zero crossings ($\alpha_i$'s can be negative).*

We will prove it by induction:

Recall, our goal is to prove the following:

**Proposition**

*If $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ is not identically zero, $f(x)$ has at most $2k-2$ zero crossings ($\alpha_i$'s can be negative).*

We will prove it by induction:

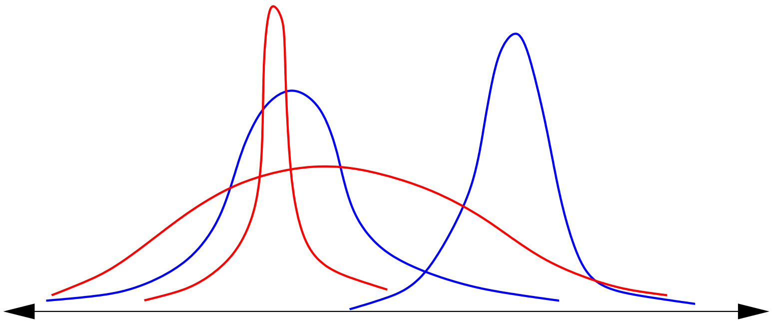
......

Start with $k = 3$ (at most 4 zero crossings),

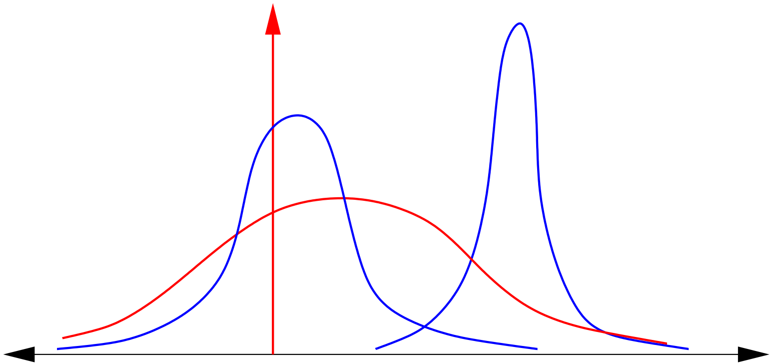
Let's prove it for $k = 4$ (at most 6 zero crossings)

......

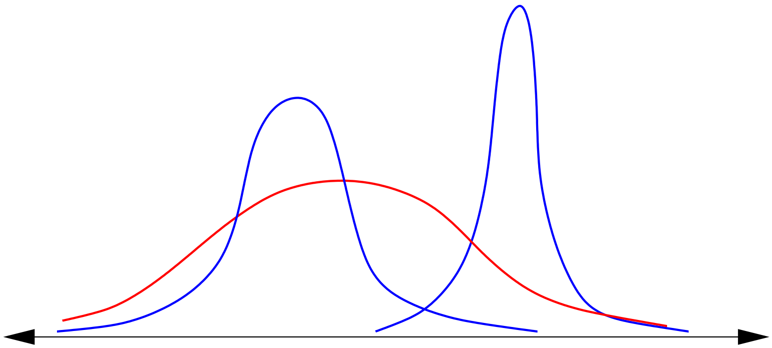
# Bounding the Number of Zero Crossings



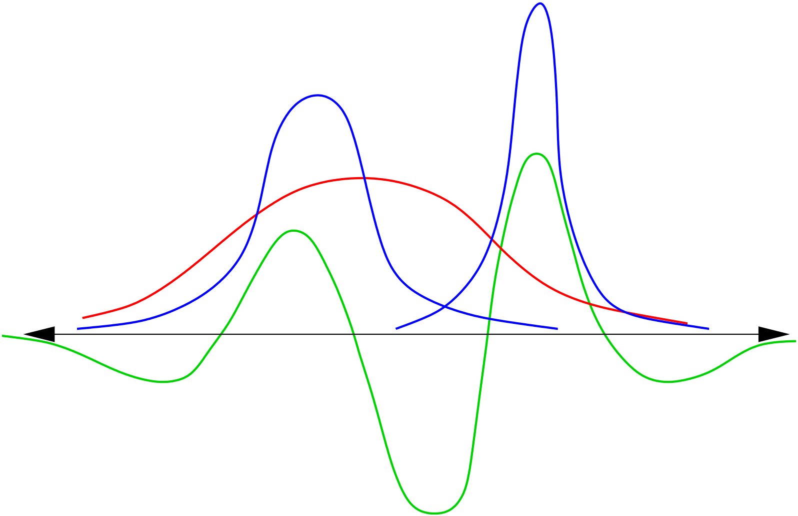
# Bounding the Number of Zero Crossings



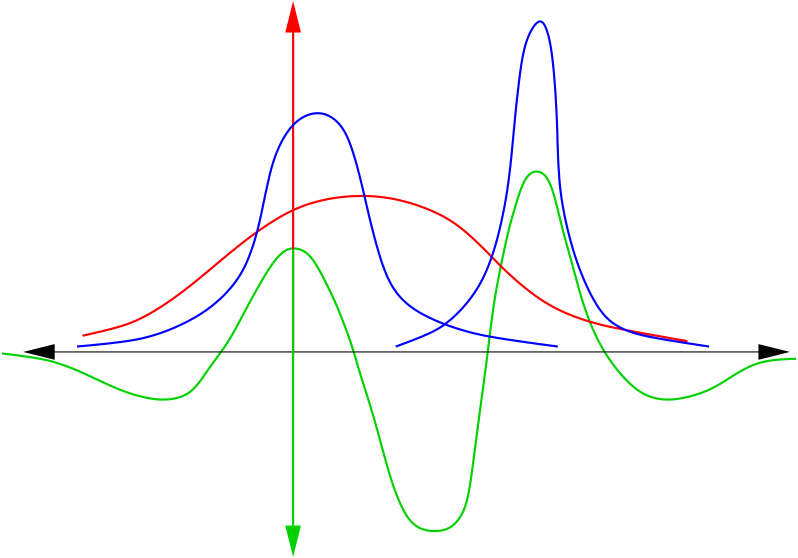
# Bounding the Number of Zero Crossings



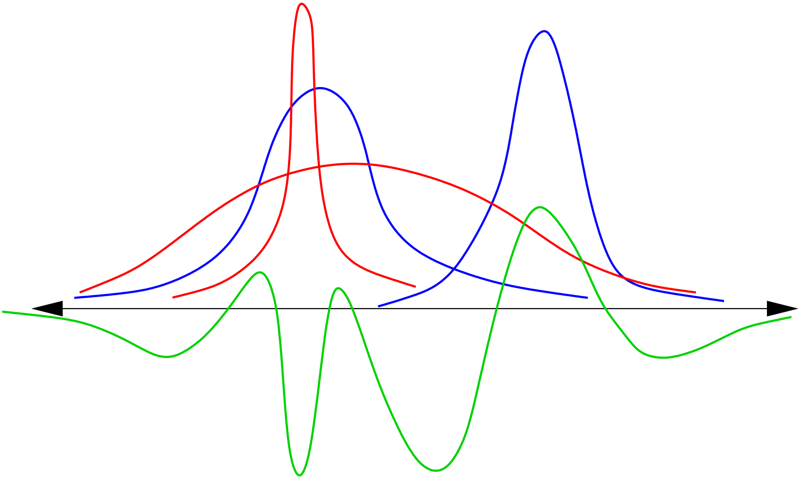
# Bounding the Number of Zero Crossings



# Bounding the Number of Zero Crossings



# Bounding the Number of Zero Crossings



Hence, the **exact** values of the first six moments determine the mixture parameters!

Hence, the **exact** values of the first six moments determine the mixture parameters!

---

Let $\Theta = \{$**valid** parameters$\}$ (in particular $w_i \in [0, 1]$, $\sigma_i \geq 0$)

Hence, the **exact** values of the first six moments determine the mixture parameters!

---

Let $\Theta = \{$**valid** parameters$\}$ (in particular $w_i \in [0, 1]$, $\sigma_i \geq 0$)

**Claim**

*Let $\theta$ be the true parameters; then the only solutions to*

$$\left\{\widehat{\theta} \in \Theta | M_r(\widehat{\theta}) = M_r(\theta) \text{ for } r = 1, 2, ...6\right\}$$

*are $(w_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and the relabeling $(1 - w_1, \mu_2, \sigma_2, \mu_1, \sigma_1)$*

Hence, the **exact** values of the first six moments determine the mixture parameters!

---

Let $\Theta = \{$**valid** parameters$\}$ (in particular $w_i \in [0,1]$, $\sigma_i \geq 0$)

**Claim**

*Let $\theta$ be the true parameters; then the only solutions to*

$$\left\{\widehat{\theta} \in \Theta | M_r(\widehat{\theta}) = M_r(\theta) \text{ for } r = 1, 2, ...6\right\}$$

*are $(w_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and the relabeling $(1 - w_1, \mu_2, \sigma_2, \mu_1, \sigma_1)$*

Are these equations stable, when we are given **noisy** estimates?

# A Type of Condition Number

Using deconvolution to isolate components, we show:

# A Type of Condition Number

Using deconvolution to isolate components, we show:

There are constants $c, C$ such that if $\epsilon < c$, the means and variances are bounded by $\frac{1}{\epsilon}$, the mixing weights are in $[\epsilon, 1-\epsilon]$ and

$$|M_r(\theta) - M_r(\widehat{\theta})| \leq \epsilon^C$$

for $r = 1, 2, \ldots 6$

# A Type of Condition Number

Using deconvolution to isolate components, we show:

There are constants $c, C$ such that if $\epsilon < c$, the means and variances are bounded by $\frac{1}{\epsilon}$, the mixing weights are in $[\epsilon, 1-\epsilon]$ and

$$|M_r(\theta) - M_r(\widehat{\theta})| \leq \epsilon^C$$

for $r = 1, 2, ...6$ then there is a permutation $\pi$ such that

$$\sum_{i=1}^{2} |w_i - \widehat{w}_{\pi(i)}| + |\mu_i - \widehat{\mu}_{\pi(i)}| + |\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon$$

# A Type of Condition Number

Using deconvolution to isolate components, we show:

There are constants $c, C$ such that if $\epsilon < c$, the means and variances are bounded by $\frac{1}{\epsilon}$, the mixing weights are in $[\epsilon, 1-\epsilon]$ and

$$|M_r(\theta) - M_r(\widehat{\theta})| \leq \epsilon^C$$

for $r = 1, 2, ...6$ then there is a permutation $\pi$ such that

$$\sum_{i=1}^{2} |w_i - \widehat{w}_{\pi(i)}| + |\mu_i - \widehat{\mu}_{\pi(i)}| + |\sigma_i^2 - \widehat{\sigma}_{\pi(i)}^2| \leq \epsilon$$

Hence, close enough estimates for the first six moments guarantee that the parameters are close too!

# A Univariate Learning Algorithm

Our algorithm:

---

---

# A Univariate Learning Algorithm

Our algorithm:

---

- Take enough samples $S$ so that $\widetilde{M}_r = \frac{1}{|S|}\sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2...6$

---

# A Univariate Learning Algorithm

Our algorithm:

---

- Take enough samples $S$ so that $\widetilde{M}_r = \frac{1}{|S|}\sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2 \ldots 6$

  (within an additive $\frac{\epsilon^C}{2}$)

---

# A Univariate Learning Algorithm

Our algorithm:

---

- Take enough samples $S$ so that $\widetilde{M}_r = \frac{1}{|S|}\sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2 \ldots 6$

  (within an additive $\frac{\epsilon^C}{2}$)

- Compute $\widehat{\theta}$ such that $M_r(\widehat{\theta})$ is close to $\widetilde{M}_r$ for $r = 1, 2 \ldots 6$

---

# A Univariate Learning Algorithm

Our algorithm:

---

- Take enough samples $S$ so that $\widetilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2 \ldots 6$

  (within an additive $\frac{\epsilon^C}{2}$)

- Compute $\widehat{\theta}$ such that $M_r(\widehat{\theta})$ is close to $\widetilde{M}_r$ for $r = 1, 2 \ldots 6$

  (within an additive $\frac{\epsilon^C}{2}$)

---

# A Univariate Learning Algorithm

Our algorithm:

---

- Take enough samples $S$ so that $\widetilde{M}_r = \frac{1}{|S|}\sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2 ... 6$

  (within an additive $\frac{\epsilon^C}{2}$)
- Compute $\widehat{\theta}$ such that $M_r(\widehat{\theta})$ is close to $\widetilde{M}_r$ for $r = 1, 2 ... 6$

  (within an additive $\frac{\epsilon^C}{2}$)

---

And $\widehat{\theta}$ must be close to $\theta$, because solutions to this system of polynomial equations are **stable**

# Summary and Discussion

- Here we gave the first efficient algorithms for learning mixtures of Gaussians with provably minimal assumptions

**Key words:** method of moments, polynomials, heat equation

# Summary and Discussion

- Here we gave the first efficient algorithms for learning mixtures of Gaussians with provably minimal assumptions

**Key words:** method of moments, polynomials, heat equation

- **Computational intractability** is everywhere in machine learning/statistics

# Summary and Discussion

- Here we gave the first efficient algorithms for learning mixtures of Gaussians with provably minimal assumptions

**Key words:** method of moments, polynomials, heat equation

- **Computational intractability** is everywhere in machine learning/statistics

- Currently, most approaches are **heuristic** and have no provable guarantees

# Summary and Discussion

- Here we gave the first efficient algorithms for learning mixtures of Gaussians with provably minimal assumptions

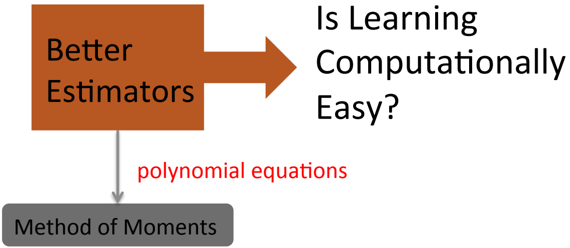
**Key words:** method of moments, polynomials, heat equation

- **Computational intractability** is everywhere in machine learning/statistics
- Currently, most approaches are **heuristic** and have no provable guarantees
- Can we design new algorithms for some of the fundamental problems in these fields?

# My Work

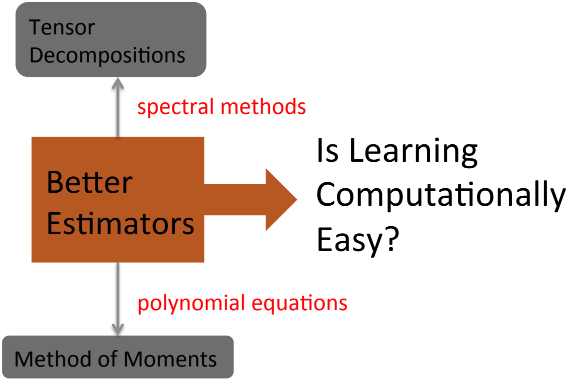
Is Learning Computationally Easy?

# My Work



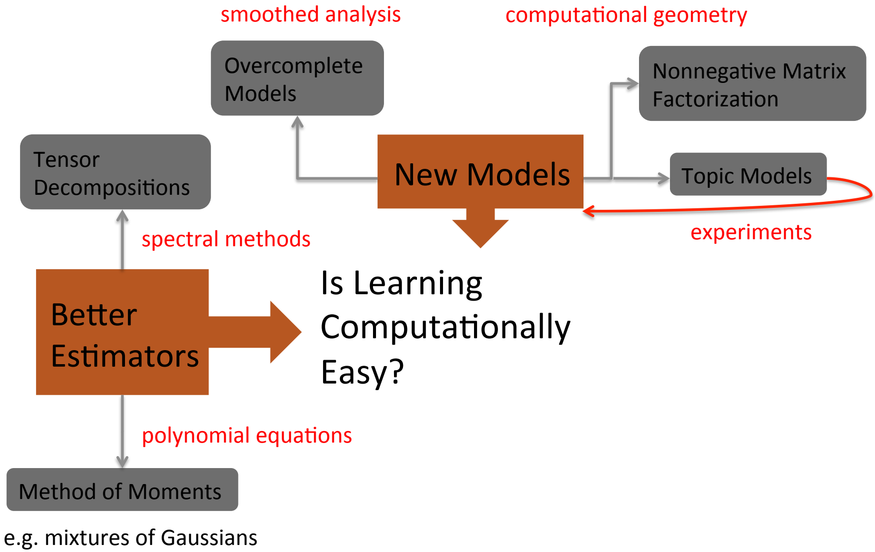

e.g. mixtures of Gaussians

# My Work

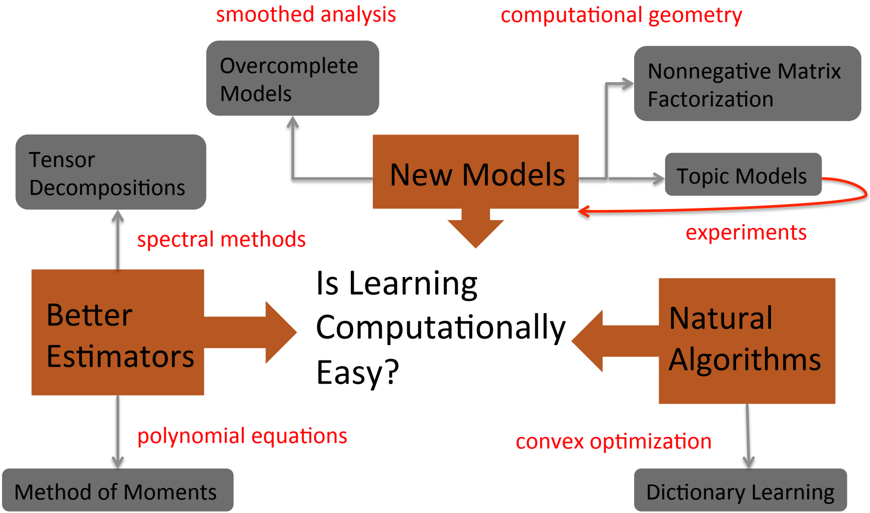



e.g. mixtures of Gaussians

# My Work


smoothed analysis
computational geometry
Overcomplete Models
Nonnegative Matrix Factorization
Tensor Decompositions
New Models
Topic Models
experiments
spectral methods
Is Learning Computationally Easy?
Better Estimators
polynomial equations
Method of Moments
e.g. mixtures of Gaussians

# My Work




e.g. mixtures of Gaussians

Thanks!