Study Group Lecture #2

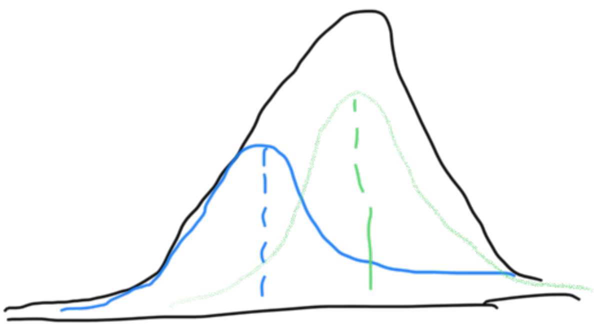
Lect #1: GMM, Compression Scheme

Today: KMV 10, MV10

$$N(\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$$GMM_2(x) = w_1 N(\mu_1, \Sigma_1; x) + (1-w_1) N(\mu_2, \Sigma_2, x)$$

captures heterogeneous populations



Given $x \sim GMM_k$

find $\vec{w}, \vec{\mu}, \vec{\Sigma} \ldots$

Pearson (1894) (crabs) 1000 (1dim)

Given $x_1, \ldots, x_m \sim GMM_2$, can we estim the 5 unknowns?

Method of Moments

$$\vec{\theta} = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2) \in \mathbb{R}^5.$$

$$\mathbb{E}_{X \sim GMM_2} X = w_1 \mu_1 + (1-w_1)\mu_2$$

$$\mathbb{E} X^2 = w_1(\mu_1^2 + \sigma_1^2) + (1-w_1)(\mu_2^2 + \sigma_2^2)$$

$$\vdots$$

$$\mathbb{E} X^r = M_r(\vec{\theta})$$

Question: Prove $M_r(\vec{\theta})$ is poly$(\vec{\theta})$

Stein Lemma $\mathbb{E}_{X \sim N(0,1)}[f'(X)] = \mathbb{E}_{X \sim N(0,1)}[X f(X)]$

$\mathbb{E}[(X-\mu)f] = \mathbb{E}[f']\sigma^2$

Pearson's 6th Moment

Find $\vec{\theta}$

$T = \{x_i\} \sim GMM_2$ Goal:

$\hat{M}_r$ = empirical r-th moment $= \frac{1}{|T|} \sum_{x_i \in T} x_i^r, \quad r = 1, \dots, 5$

Solve $\{\underbrace{\hat{M}_r}_{\in \mathbb{R}} = M_r(\vec{\theta})\}_{r \in [5]} \to \vec{\theta}_1 \in \mathbb{R}^5, \vec{\theta}_2 \in \mathbb{R}^5$

verify which is closer in 6th moment.

dim #mixt

$d=1, k=2$ (1894) $\to$ 6

(a) $d=1, k>2$???

(b) Why? Distr $\Leftrightarrow$ Moments

Moment Problems (1890++)

Given $(m_n)_{n \geq 0}$: $m_n = \int_{\mathcal{X}} x^n \, d\mu$, does there exist pos. Borel measure μ?

If so, is it unique?

μ has finite support $\overset{\text{by}}{\Rightarrow}$ Linear algebra.

μ has compact support $\overset{\text{by}}{\Rightarrow}$ Weierstrass thm & Riesz repr.

$\mathcal{X} = [0,1]$ (1923, Hausdorff)

$m_n = \int_0^1 x^n \, d\mu$

has $\boxed{\text{unique sol'n}}$ (determinate)

$\Leftrightarrow (m_n)$ completely monotonic

$(\partial^k m)_n = \sum_{i=0}^{k} \binom{k}{i} (-1)^i m_{n+i}$

$(-1)^{k} (\Delta^{k} m)_n \geq 0 \quad \forall n, k \geq 0$

$X = \mathbb{R}$ (1920)

Hamburger moments (H.M.)

$m_n = \int_{-\infty}^{\infty} x^n \, d\mu$

Hankel matrix

(H.M) is solvable $\Leftarrow$ $A = \begin{pmatrix} m_0 & m_1 & m_2 & \cdots \\ m_1 & m_2 & m_3 & \cdots \\ m_2 & m_3 & m_4 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \succeq 0$

$\boxed{\text{solution space is convex}}$

$K(x,y) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$

kernel.

$\forall$ seq (v_j) with finite support

$$v A v^{*} \geq 0$$

sketch

"only if" : Say μ exists.

$$\sum_{j,k} m_{j+k} v_j \overline{v_k} = \int_{-\infty}^{\infty} \Big| \sum_{j \geq 0} v_j x^j \Big|^2 d\mu \geq 0$$

$\Leftarrow$ ...

Carleman's condition (1926)

(Suff. condition) $\boxed{\mu \text{ on } \mathbb{R}}$

If $\sum_{n=1}^{\infty} m_{2n}^{-\frac{1}{2n}} = +\infty$ then (quasianalytic)

(m_n) is determinate

---

dim

$d=1$, $k=2$ (6 "matched" moments $\Rightarrow$ "Learning")

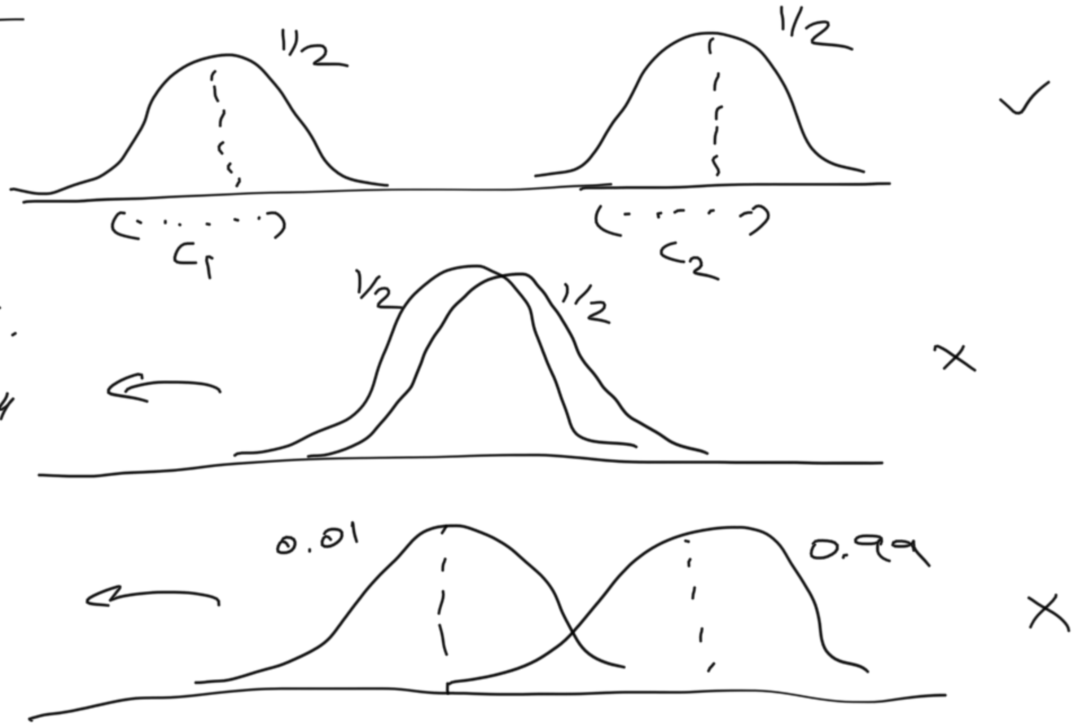
$d=1, k=?$

Is Pearson "Robustly Identifiable"

Part 2

"distr. gap"

"weights gap"



① Clustering $\Rightarrow$ Strong Separability ~~Assumptions~~.

Today $\Rightarrow$ NO

② GMM is identifiable (Teicher) g: param $\to$ density "1-1"

③ $X \sim N(\mu, \Sigma)$ ✓

$v^T X$ ✓

Proj of GMM is one-dim GMM.

---

k upper bound

Output: $\hat{F} = \sum_{i=1}^{\hat{k}} \hat{w}_i \hat{F}_i$

correctly learn $\hat{k}$

(v1) & approx (w_i, μ_i, Σ_i) $\longrightarrow$

(v2) $\to$ $TV(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| dx$ $\longrightarrow$
$\in [0,1]$

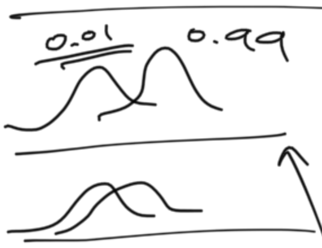
Plan

$\boxed{\text{algo 1}}$ (1d) k

$\boxed{\text{algo 2}}$ (high d)

$\boxed{\text{algo 3}}$

$TV(\cdot,\cdot) < \varepsilon \Rightarrow \|\cdot\|_2 , \|\cdot\|_F < O(\varepsilon)$



How hard ?

Inspiration from Lin. Alg

"$\kappa(A)$" $= \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1$

$\kappa \to \infty$ bad

$A\underline{x} = b$

robust

stable

Condition #. (GMM)

$$\kappa(F) = \frac{1}{\min\{ \{w_i\}_i , TV(F_i, F_j)_{ij} \}}$$

$[0,1]$ ✓

Any algo requires $\Omega(\kappa(F))$ samples.

Main Result (Algo 3) (2010)

$\forall k \geq 1$, $\exists\ C_k = C(k)$ s.t.

$\forall d$ GMM F, $\delta > 0$ $\qquad n > \left(\frac{\kappa(F) \cdot d}{\varepsilon \delta} \right)^{C_k}$

the estim. algo 3 finds $\hat{F}$ s.t.

w.p $\geq 1 - \delta$:

- $\hat{k} = k$ $\qquad \exists\ \pi \in S_n$
- $|w_i - \hat{w}_{\pi(i)}| \leq \varepsilon$ $\qquad \forall i \in [k]$
  $TV(F_i, \hat{F}_{\pi(i)}) \leq \varepsilon$

Runtime: poly(n)

Pearson was right! (Kalai, Moitra, G Valiant) STOC 2010

$d = 1, k \to \boxed{4k-2}$

Input: $\{x_i\} \sim$ 1d GMM_k , ε

compute $4k-2$ for (w_i, μ_i, σ_i^2) :

grid search $\left\{ \begin{array}{l} w_i \in [0,1] \\ \mu_i \in [\min x_j, \max x_j] \\ \sigma_i \in [0, \max x_j] \\ \text{grid } \epsilon^{C_k} \end{array} \right.$

Return param set where first $4k-2$ ....

Part 3: Robust Identifiability (Finite sample)

TV is too strong sometimes

$TV\left(\frac{S_n}{\sqrt{n}}, N\right) = 1 \ \forall n$

Thm (Wanted)

F, F' GMM $< k$ comp.

with $\kappa(F), \kappa(F') < \frac{1}{\varepsilon}$

If $\forall i \leq \boxed{4k-2}$: $|M_i(F) - M_i(F')| \leq \varepsilon^{C_k}$

$\Rightarrow$ (1) F, F' same $\hat{k}$

(2) param gap $\leq \varepsilon$ ✓

$\|\mu - \mu'\|_2$, $\|\Sigma - \Sigma'\|_F$

Observations:

1. If two distr (bounded) have $\downarrow$ TV

f, g

$\Rightarrow$ Low-order moments must be close.

$\not\Leftarrow$ $h = f - g$

$4k-2$

2. Dream 1: Algo 2

Low-ord moment "same" $\Rightarrow$ params "same"

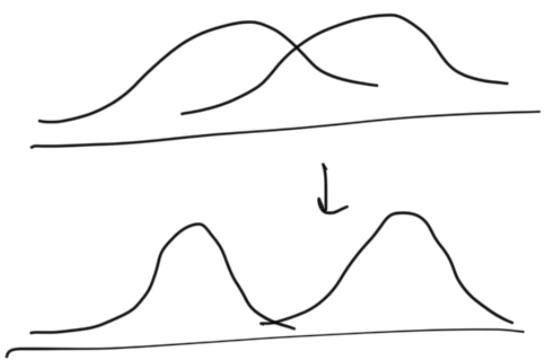
Dream 2: Generally:

params same. ~~X0~~ TV small (Algo 3)

Deconvolution & Moments

zero crossings dictates # moments required.

GMM_k: $f_1 - f_2$



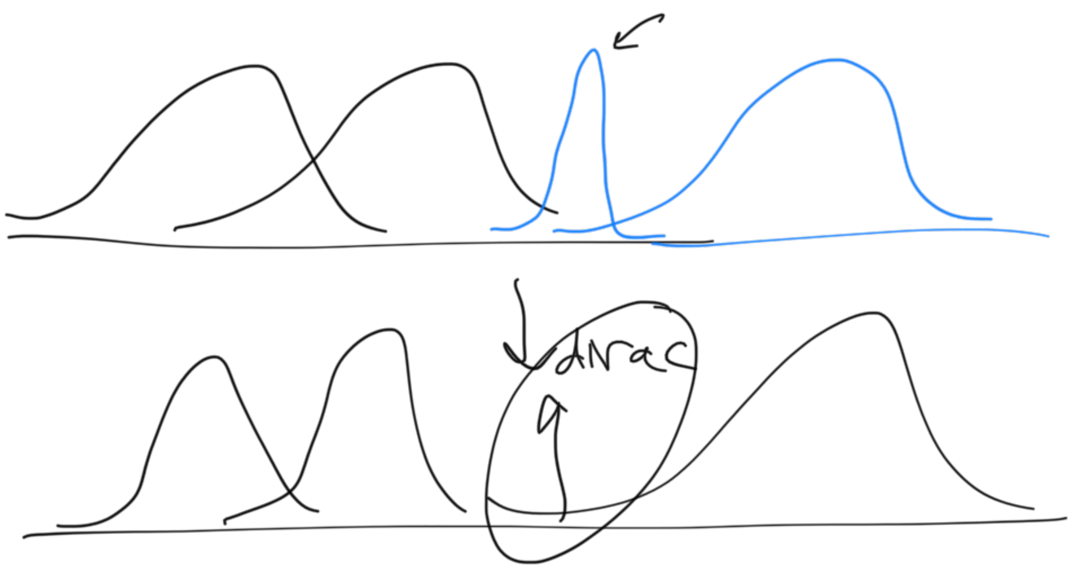
α-deconvolve

$N(\mu, \sigma^2) \to N(\mu, \sigma^2 - \alpha)$

Convolution $\longrightarrow$ heat equation

$f * g \xrightarrow{\mathcal{F}} \hat{f} \cdot \hat{g}$

Deconvolution is heat equation reversed.



Let F, F' are ε-standard

"param gap"

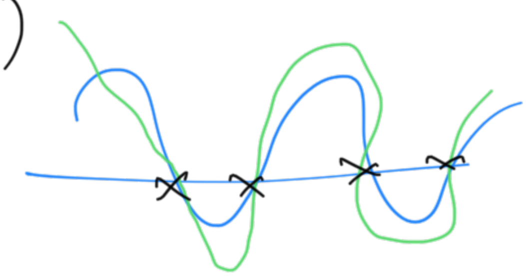
~~Sketch~~

1. Let F, F' $TV(F, F') > \varepsilon$

$$f(x) = F(x) - F'(x)$$

$$\int |f(x)| > 0$$

2. f has at most 6 $\boxed{4k-2}$ crossings (? thm)



3. ∃ poly deg 6 $\boxed{4k-2}$ whose sign matches f.

$$p(x) \to \sum c_i x^i$$

$$0 \overset{TV}{<} \left| \int f(x) p(x) \, dx \right| =$$

$$= \left| \int \sum_{i=1}^{6} c_i x^i f(x) \, dx \right|$$

$$\leq \sum_i |c_i| \left| \int x^i f(x) \, dx \right|$$

$$\leq \sum_i |c_i| \left| M_i(F) - M_i(F') \right|$$

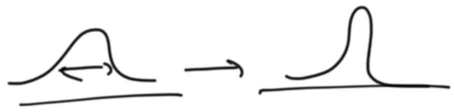
∃ i that Moments have gap ?

Steps for main thm ← (1d)

F, F' ε-standard) → Moment gap (param gap?)) $\Rightarrow$ ... gap.

"
$$\min_{\pi} \sum_{i=1}^{k} \{ |w_i - w'_{\pi(i)}| + \|\mu_i - \mu'_{\pi(i)}\|_2 + \|\Sigma_i - \Sigma'_{\pi(i)}\|_F \} \geq \varepsilon,$$

(1) Strip away common parts.



(2) $\boxed{\text{Deconvolve}}$ the min variance part

(3) This part $\propto$ Dirac $\boxed{\mathcal{F}_\alpha}$

$\Downarrow$

generates a TV dist. gap in $\mathcal{F}_\alpha$-space

(4) TV gap in $\mathcal{F}_a$ $\Rightarrow$ Low-order moment gap (in $\mathcal{F}_a$ space)

$\exists i : M_i(\mathcal{F}_a(\mathcal{F})) - M_i(\mathcal{F}_\alpha(\mathcal{F}')) \geq \varepsilon$

(5) Magic Lemma:

Deconvolution preserves low-order moments.

"If, in $\mathcal{F}_\alpha$ space, has gap

$\Rightarrow$ original space has gap

$\boxed{\text{L}}$

---

why? $f(x): \mathbb{R} \to \mathbb{R}$ analytic & has n zeros

$\Rightarrow g(x) = f(x) * N(0, \sigma^2, x)$

has at most n zeros

TEN

$$\sum^{r} | M_i(\mathcal{F}_\alpha(\mathcal{F})) - M_i(\mathcal{F}_\alpha(\mathcal{F}')) | <$$

$$\leq \frac{(r+1)!}{\lfloor r/2 \rfloor !} \left| \sum_{i=1}^{r} M_i(F) - M_i(F') \right|$$

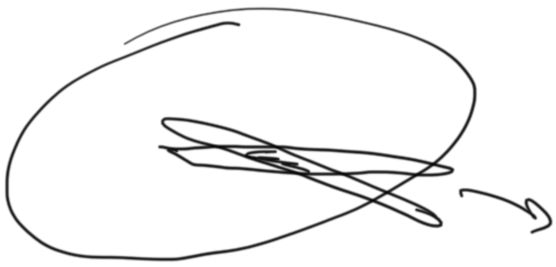
$\boxed{\text{Algo 2}}$

high-dim

d^2

Algo 3

param close $\not\Rightarrow$ TV $<< \varepsilon$



TV = 1

measure 0

Lect #3: Mao, Opt. Trans.

#4 : Representation theory

#5 : DP

#6 : DP2

#7 : LDP