Estimating Mean Functionals in the Gaussian Vector Model

Nicolas Verzelen

INRA Montpellier

Luminy, December 2017

Setting

Gaussian Vector Model

$$Y \sim \mathcal{N}(\theta, I_n) \ , \qquad \text{where } \theta \text{ is unknown}$$

Functional estimation : Fix a function $f : \mathbb{R}^n \to \mathbb{R}$.
We want to estimate $f(\theta)$ based on Y.

Examples :

- $L(\theta) = \sum_{i=1}^n \theta_i$ (linear functional)
- $Q(\theta) = \sum_{i=1}^n \theta_i^2$ (quadratic functional)
- $m(\theta) = \min_{i=1}^n \theta_i$ (minimum functional)

Setting

Gaussian Vector Model

$$Y \sim \mathcal{N}(\theta, I_n) \ , \qquad \text{where } \theta \text{ is unknown}$$

Functional estimation : Fix a function $f : \mathbb{R}^n \to \mathbb{R}$.
We want to estimate $f(\theta)$ based on Y.

Examples :

- $L(\theta) = \sum_{i=1}^n \theta_i$ (linear functional)
- $m(\theta) = \min_{i=1}^n \theta_i$ (minimum functional)

Objectives

Structural Assumption : θ is sparse in some sense
$\rightsquigarrow$ at most k of the components of θ differ from some basal value θ_0.

This talk :

- Characterizing minimax adaptive risks of $L(\theta)$ and $m(\theta)$.
- Focusing on some proof techniques for the lower bounds

Objectives

Structural Assumption : θ is sparse in some sense
$\rightsquigarrow$ at most k of the components of θ differ from some basal value θ_0.

This talk :

- Characterizing minimax adaptive risks of $L(\theta)$ and $m(\theta)$.
- Focusing on some proof techniques for the lower bounds
- . . . acknowledging some contributions of Oleg and Sasha

Based on two joint work with

- **O. Collier, L. Comminges, and A. Tsybakov**. Optimal Adaptive Estimation of Linear Functional Under Sparsity. `arXiv:1611.09744` [AoS(+18)]
- **A. Carpentier, S. Delattre, and E. Roquain**. Minimax estimation of the shift of a random vector with application to multiple testing. `arXiv:1801.????`

Linear Functional Estimation problem

Problem : Estimating $L(\theta) = \sum_{i=1}^{n} \theta_i$

Structural Assumption : At most k components of θ are non zero.
$\rightsquigarrow \theta \in \Theta_0[k] = \{\theta, \|\theta\|_0 \leq k\}$.

Linear Functional Estimation problem

Problem : Estimating $L(\theta) = \sum_{i=1}^{n} \theta_i$

Structural Assumption : At most k components of θ are non zero.
$\rightsquigarrow \theta \in \Theta_0[k] = \{\theta, \|\theta\|_0 \leq k\}$.

Maximal Risk over $\Theta_0[k]$: $\quad \Psi_{k,n}(\widehat{L}) = \sup_{\theta \in \Theta_0[k]} \mathbb{E}_\theta[\widehat{L} - L(\theta)]^2$

Minimax Risk over $\Theta_0[k]$: $\quad \Psi^*_{k,n} = \inf_{\widehat{L}} \sup_{\theta \in \Theta_0[k]} \mathbb{E}_\theta[\widehat{L} - L(\theta)]^2$

Linear Functional Estimation problem

Problem : Estimating $L(\theta) = \sum_{i=1}^{n} \theta_i$

Structural Assumption : At most k components of θ are non zero.
$\rightsquigarrow \theta \in \Theta_0[k] = \{\theta, \|\theta\|_0 \leq k\}$.

Maximal Risk over $\Theta_0[k]$: $\Psi_{k,n}(\widehat{L}) = \sup_{\theta \in \Theta_0[k]} \mathbb{E}_\theta[\widehat{L} - L(\theta)]^2$

Minimax Risk over $\Theta_0[k]$: $\Psi^*_{k,n} = \inf_{\widehat{L}} \sup_{\theta \in \Theta_0[k]} \mathbb{E}_\theta[\widehat{L} - L(\theta)]^2$

Goal 1 : For each $k \in [n]$, computing $\Psi^*_{k,n}$.

Goal 2 : [Adaptation] Building $\widetilde{L}$ performing as best as possible simultaneously over all $\Theta_0[k]$.

Related work : see e.g. Ibragimov and Hasminskii('84) ; Klemela and Tsybakov('01) Cai and Low('04,'05) ; Golubev and Levit('04) ; Collier et al.('17) ; Collier and Dalalyan (today)...

(Non Adaptive) Minimax Risk

Heuristic : When k is small, threshold the small of components of Y_i.

Define the estimator

$$\widehat{L}_k = \begin{cases} \sum_{j=1}^n Y_j \mathbf{1}\{Y_j^2 > 2\log(1+n/k^2)\}, & \text{if } k < \sqrt{n}, \\ \sum_{j=1}^n Y_j, & \text{otherwise,} \end{cases}$$

Theorem (Collier et al.('17))

For $k = 1, \ldots, n$,

$$\Psi_{k,n}[\widehat{L}_k] \lesssim k^2 \log\left(1 + \frac{n}{k^2}\right) \asymp \Psi^*_{k,n}$$

Remark :

- For $k < \sqrt{n}/2$, $\Psi^*_{k,n} \asymp k^2 \log\left(\frac{n}{k^2}\right)$
- For $k \geq \sqrt{n}/2$, $\Psi^*_{k,n} \asymp n$

Step 1. From Estimation to Hypothesis testing.

- μ_0 : prior on $\Theta_0[k]$, such that $L(\theta) = 0$, μ_0 a.s.
- μ_1 : prior on $\Theta_0[k]$, such that $L(\theta) = T$, μ_1 a.s.

Mixture distribution : $\mathbf{P}_0 = \int \mathbb{P}_\theta \, d\mu_0(\theta)$, $\mathbf{P}_1 = \int \mathbb{P}_\theta \, d\mu_1(\theta)$.

$$
\begin{aligned}
\sup_{\theta \in \Theta_0[k]} \mathbb{E}[\widehat{L} - L(\theta)]^2 &\geq \frac{1}{2}\Big[\mathbf{E}_0[\widehat{L}^2] + \mathbf{E}_1[\widehat{L} - T]^2\Big] \\
&\geq \frac{T^2}{8}\Big[\mathbf{P}_0[\widehat{L} > T/2] + \mathbf{P}_0[\widehat{L} \leq T/2]\Big] \\
&\geq \frac{T^2}{8}\Big[1 - \|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}\Big]
\end{aligned}
$$

Lower Bound : Vanilla Le Cam's two points Method

Step 1. From Estimation to Hypothesis testing.

- μ_0 : prior on $\Theta_0[k]$, such that $L(\theta) = 0$, μ_0 a.s.
- μ_1 : prior on $\Theta_0[k]$, such that $L(\theta) = T$, μ_1 a.s.

Mixture distribution : $\mathbf{P}_0 = \int \mathbb{P}_\theta \, d\mu_0(\theta)$, $\mathbf{P}_1 = \int \mathbb{P}_\theta \, d\mu_1(\theta)$.

$$\begin{aligned}
\sup_{\theta \in \Theta_0[k]} \mathbb{E}[\widehat{L} - L(\theta)]^2 &\geq \frac{1}{2}\Big[\mathbf{E}_0[\widehat{L}^2] + \mathbf{E}_1[\widehat{L} - T]^2\Big] \\
&\geq \frac{T^2}{8}\Big[\mathbf{P}_0[\widehat{L} > T/2] + \mathbf{P}_0[\widehat{L} \leq T/2]\Big] \\
&\geq \frac{T^2}{8}\Big[1 - \|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}\Big]
\end{aligned}$$

Step 2. Bound of $\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}$.
Distance between two mixtures difficult to control. $\rightsquigarrow$ Take $\mu_0 = \delta_0$ $(\mathbf{P}_0 = \mathbb{P}_0)$.

$$\| \mathbb{P}_0 - \mathbf{P}_1\|_{TV}^2 \leq \chi^2(\mathbb{P}_0; \mathbf{P}_1) = \mathbb{E}_0\left[\Big(\frac{d\mathbf{P}_1}{d\mathbb{P}_0}\Big)^2\right] - 1$$

explicit computations + careful choice of T and μ_1 $\Rightarrow$ Desired lower bound

An adaptive procedure ?

$\rightsquigarrow$ Lepski-type method to build an adaptive procedure.

Introduce non-adaptive estimators $\widetilde{L}_k$ indexed by $k = 1, \ldots, n$:

$$\widetilde{L}_k = \begin{cases} \sum_{j=1}^n Y_j \mathbf{1}\{Y_j^2 > \alpha \log\left(1 + \frac{n \log n}{k^2}\right)\}, & \text{if } k \leq \sqrt{n \log n/2}, \\ \sum_{j=1}^n Y_j, & \text{otherwise,} \end{cases}$$

An adaptive procedure ?

$\rightsquigarrow$ Lepski-type method to build an adaptive procedure.

Introduce non-adaptive estimators $\widetilde{L}_k$ indexed by $k = 1, \ldots, n$:

$$\widetilde{L}_k = \begin{cases} \sum_{j=1}^n Y_j \mathbf{1}\{Y_j^2 > \alpha \log\left(1 + \frac{n \log n}{k^2}\right)\}, & \text{if } k \leq \sqrt{n \log n / 2}, \\ \sum_{j=1}^n Y_j, & \text{otherwise}, \end{cases}$$

Take $\widetilde{L} \triangleq \widetilde{L}_{\hat{k}}$ where

$$\hat{k} \triangleq \min\left\{k \in \{1, \ldots, \lfloor\sqrt{n \log n}\rfloor\} : |\tilde{L}_k - \tilde{L}_{k'}|^2 \leq \beta k'^2 \log\left(1 + \frac{n \log n}{k'^2}\right), \text{ for all } k' > k\right\}.$$

Theorem (Collier et al.(18))

For $k = 1, \ldots, n$ and any $\theta \in \Theta_0[k]$,

$$\sup_{\theta \in \Theta_0[k]} \mathbb{E}_\theta[(\widetilde{L} - L(\theta))^2] \lesssim \Phi_{k,n} \triangleq k^2 \log\left(1 + \frac{n \log(n)}{k^2}\right)$$

Higher than $\Psi^*_{k,n} \asymp k^2 \log\left(1 + \frac{n}{k^2}\right)$ for $k \geq \sqrt{n / \log(n)}$

Adaptation : characterization and lower bound

(Following Tsybakov('98)), a function $k \mapsto \Phi_{k,n}$ is *an adaptive rate of convergence* if

(i) There exists an estimator $\widehat{L}$ such that, for all k,

$$\max_{k=1,\dots,n} \sup_{\theta\in\Theta_0[k]} \mathbb{E}_\theta(\hat{L} - L(\theta))^2/\Phi_{k,n} \lesssim 1$$

(ii) And for all functions $k \mapsto \Phi'_{k,n}$ satisfying (i),

$$\min_k \frac{\Phi'_{k,n}}{\Phi_{k,n}} \to 0 \qquad \Rightarrow \qquad \max_k \frac{\Phi'_{k,n}}{\Phi_{k,n}} \cdot \min_k \frac{\Phi'_{k,n}}{\Phi_{k,n}} \to \infty \ .$$

Adaptation : characterization and lower bound

(Following Tsybakov('98)), a function $k \mapsto \Phi_{k,n}$ is *an adaptive rate of convergence* if

(i) There exists an estimator $\widehat{L}$ such that, for all k,

$$\max_{k=1,\dots,n} \sup_{\theta \in \Theta_0[k]} \mathbb{E}_\theta (\hat{L} - L(\theta))^2 / \Phi_{k,n} \lesssim 1$$

(ii) And for all functions $k \mapsto \Phi'_{k,n}$ satisfying (i),

$$\min_k \frac{\Phi'_{k,n}}{\Phi_{k,n}} \to 0 \qquad \Rightarrow \qquad \max_k \frac{\Phi'_{k,n}}{\Phi_{k,n}} \cdot \min_k \frac{\Phi'_{k,n}}{\Phi_{k,n}} \to \infty \; .$$

Theorem (Collier et al.('18))

Any estimator $\widehat{L}$ that satisfies

$$\sup_{\theta \in \Theta_0[k]} \mathbf{E}_\theta \left[\left(\widehat{L} - L(\theta) \right)^2 \right] \leq c \Phi_{k,n} \quad \text{for some } k \geq n^{1/4}$$

has a degenerate maximal risk over $\Theta_0[1]$, that is

$$\sup_{\theta \in \Theta_0[1]} \mathbf{E}_\theta \left[\left(\widehat{L} - L(\theta) \right)^2 \right] \gtrsim n^{1/4} \; .$$

Proof Sketch : Asymmetric Two-point Method

Lemma

For $k \geq n^{1/4}$,

$$R(k) \triangleq \inf_{\tilde{L}} \left\{ \frac{\mathbb{E}_0(\tilde{L} - L(0))^2}{n^{1/4}} + \sup_{\theta \in \Theta_0[k]} \frac{\mathbb{E}_\theta(\tilde{L} - L(\theta))^2}{\Phi_{k,n}} \right\} \gtrsim 1$$

Proof Sketch : Asymmetric Two-point Method

Lemma

For $k \geq n^{1/4}$,

$$R(k) \triangleq \inf_{\tilde{L}} \left\{ \frac{\mathbb{E}_0(\tilde{L} - L(0))^2}{n^{1/4}} + \sup_{\theta \in \Theta_0[k]} \frac{\mathbb{E}_\theta(\tilde{L} - L(\theta))^2}{\Phi_{k,n}} \right\} \gtrsim 1$$

Proof : Build μ_1 on $\Theta_0[k]$ s.t. $L(\theta) \asymp \Phi_{k,n}^{1/2}$, μ_1 a.s.

$$\begin{aligned} R(k) \quad &\geq \quad \inf_{\tilde{L}} \left\{ \frac{\mathbb{E}_0 \, \tilde{L}^2}{n^{1/4}} + \frac{\mathbf{E}_1(\tilde{L} - L)^2}{\Phi_{k,n}} \right\} \\ &\gtrsim \quad \inf_{\mathcal{A}} \left[\mathbb{P}_0(\mathcal{A}^c) \frac{\Phi_{k,n}}{n^{1/4}} + \mathbf{P}_1(\mathcal{A}) \right] \end{aligned}$$

Proof Sketch : Asymmetric Two-point Method

Lemma

For $k \geq n^{1/4}$*,*

$$R(k) \triangleq \inf_{\tilde{L}} \left\{ \frac{\mathbb{E}_0(\tilde{L} - L(0))^2}{n^{1/4}} + \sup_{\theta \in \Theta_0[k]} \frac{\mathbb{E}_\theta(\tilde{L} - L(\theta))^2}{\Phi_{k,n}} \right\} \gtrsim 1$$

Proof : Build μ_1 on $\Theta_0[k]$ s.t. $L(\theta) \asymp \Phi_{k,n}^{1/2}$, μ_1 a.s.

$$\begin{aligned} R(k) \quad &\geq \quad \inf_{\tilde{L}} \left\{ \frac{\mathbb{E}_0\, \tilde{L}^2}{n^{1/4}} + \frac{\mathbf{E}_1(\tilde{L} - L)^2}{\Phi_{k,n}} \right\} \\ &\gtrsim \quad \inf_{\mathcal{A}} \left[\mathbb{P}_0(\mathcal{A}^c) \frac{\Phi_{k,n}}{n^{1/4}} + \mathbf{P}_1(\mathcal{A}) \right] \end{aligned}$$

Lemma

Let P *and* Q *be two probability measures. For any* $q > 0$*,*

$$\inf_{\mathcal{A}} \{P(\mathcal{A})q + Q(\mathcal{A}^c)\} \geq \frac{1}{2}\big(1 - \frac{1}{q}(\chi^2(Q,P) + 1)\big).$$

Setting and Motivation

Problem : Estimating $m(\theta) \triangleq \min_{i=1,\dots,n} \theta_i$.

Structural Assumption : At most k components of θ are larger than $m(\theta)$.
$\rightsquigarrow \Theta_m[k] = \big\{\theta,\ \sum_{i=1}^n \mathbf{1}\big\{\theta_i > m(\theta)\big\} \leq k\big\}$.

Setting and Motivation

Problem : Estimating $m(\theta) \triangleq \min_{i=1,\dots,n} \theta_i$.

Structural Assumption : At most k components of θ are larger than $m(\theta)$.
$\rightsquigarrow \Theta_m[k] = \big\{\theta,\ \sum_{i=1}^n \mathbf{1}\big\{\theta_i > m(\theta)\big\} \leq k\big\}$.

Two motivations :

- Multiple Testing and FDR control with equicorrelation $Y \sim \mathcal{N}(\mu, aI_n + bJ_n)$
  Testing $H_{0,i} : \mu_i = 0$ vs $H_{1,i} : \mu_i > 0$
  $\rightsquigarrow$ Factor model $Y = \gamma + \mu_i + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, a)$.
  $\rightsquigarrow$ Estimating $\gamma = m(\theta)$ = removing the unknown factor .

Problem : Estimating $m(\theta) \triangleq \min_{i=1,\dots,n} \theta_i$.

Structural Assumption : At most k components of θ are larger than $m(\theta)$.
$\rightsquigarrow \Theta_m[k] = \big\{\theta,\ \sum_{i=1}^n \mathbf{1}\big\{\theta_i > m(\theta)\big\} \leq k\big\}$.

Two motivations :

- Multiple Testing and FDR control with equicorrelation $Y \sim \mathcal{N}(\mu, aI_n + bJ_n)$
  Testing $H_{0,i} : \mu_i = 0$ vs $H_{1,i} : \mu_i > 0$
  $\rightsquigarrow$ Factor model $Y = \gamma + \mu_i + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, a)$.
  $\rightsquigarrow$ Estimating $\gamma = m(\theta) =$ removing the unknown factor .

- Mean Estimation in the presence of a one-sided contamination
  For at least $n - k$ observations, $Y_i \sim \mathcal{N}(m(\theta), 1)$
  At most k contaminated observations, with $Y_i \sim \mathcal{N}(\theta_i, 1)$ and $\theta_i > m(\theta)$.
  $\neq$ Huber-contamination model (contamination is one-sided and normal)

Minimax Risk over $\Theta_m[k]$: $\inf_{\widehat{m}} \sup_{\theta \in \Theta_m[k]} \mathbb{E}_\theta \left|\widehat{m} - m(\theta)\right|$

Preliminary Observations

Empirical Median : $\widehat{m}_0 = \mathrm{Med}(Y)$.

Proposition

$$\sup_{\theta \in \Theta_m[k]} \mathbb{E}_\theta \left|\widehat{m} - m(\theta)\right| \lesssim \frac{k}{n} + \frac{1}{\sqrt{n}}, \qquad \textit{if } k \leq n(1/2 - \kappa)$$

$\rightsquigarrow$ This risk is optimal for $k \leq \sqrt{n}$.

Preliminary Observations

Empirical Median : $\widehat{m}_0 = \mathrm{Med}(Y)$.

Proposition

$$\sup_{\theta \in \Theta_m[k]} \mathbb{E}_\theta \left| \widehat{m} - m(\theta) \right| \lesssim \frac{k}{n} + \frac{1}{\sqrt{n}}, \qquad \textit{if } k \leq n(1/2 - \kappa)$$

$\rightsquigarrow$ This risk is optimal for $k \leq \sqrt{n}$.

Empirical Minimum : $\widehat{m}_\infty = \min_{i=1,\dots,n}(Y_i)$.

Proposition

$$\sup_{\theta \in \Theta_m[k]} \mathbb{E}_\theta \left| \widehat{m} - m(\theta) \right| \lesssim \sqrt{\log(n)}$$

Question : To what extent are these rates optimal ?

Towards a Minimax Lower Bound : Moment Matching

Two-point reduction to $\mathbb{P}_0$ and $\mathbf{P}_1 = \int \mathbb{P}_\theta \, d\mu_1(\theta)$ leads to suboptimal rate.

Need to consider composite-composite reduction of $\mathbf{P}_0 = \int \mathbb{P}_\theta \, d\mu_0(\theta)$ vs $\mathbf{P}_1$

Difficulty : $\chi^2(\mathbf{P}_0; \mathbf{P}_1) = \int \left(\frac{d\mathbf{P}_1}{d\mathbf{P}_0}\right)^2 d\mathbf{P}_0 - 1$ is not tractable

Towards a Minimax Lower Bound : Moment Matching

Two-point reduction to $\mathbb{P}_0$ and $\mathbf{P}_1 = \int \mathbb{P}_\theta \, d\mu_1(\theta)$ leads to suboptimal rate.

Need to consider composite-composite reduction of $\mathbf{P}_0 = \int \mathbb{P}_\theta \, d\mu_0(\theta)$ vs $\mathbf{P}_1$

Difficulty : $\chi^2(\mathbf{P}_0; \mathbf{P}_1) = \int \left(\frac{d\mathbf{P}_1}{d\mathbf{P}_0}\right)^2 d\mathbf{P}_0 - 1$ is not tractable

Workaround : Relating $\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}$ to measures of proximities between μ_0 and μ_1 :

Towards a Minimax Lower Bound : Moment Matching

Two-point reduction to $\mathbb{P}_0$ and $\mathbf{P}_1 = \int \mathbb{P}_\theta \, d\mu_1(\theta)$ leads to suboptimal rate.

Need to consider composite-composite reduction of $\mathbf{P}_0 = \int \mathbb{P}_\theta \, d\mu_0(\theta)$ vs $\mathbf{P}_1$

Difficulty : $\chi^2(\mathbf{P}_0; \mathbf{P}_1) = \int \left(\frac{d\mathbf{P}_1}{d\mathbf{P}_0}\right)^2 d\mathbf{P}_0 - 1$ is not tractable

Workaround : Relating $\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}$ to measures of proximities between μ_0 and μ_1 :

Two (three) fruitful approaches :

- (densities of $\mathbf{P}_0$ and $\mathbf{P}_1$ are matching on a wide set) (e.g Chen et al.('15))
- Fourier Transforms of μ_0 and μ_1 are matching on a wide interval. (e.g. Moitra and Valiant('10) ; Cai and Jin('10), Carpentier and V.('17),. . . ,)
- First moments of μ_0 and μ_1 are matching. (Lepski, Nemirovski and Spokoiny('99) for L_r norm).
  See also : Cai and Low('11) (L_1 norm) ; Jiao et al.('15) ; Wu and Yang('16) ; (Entropy Estimation) Bandeira et al.('17) (Multi-reference Alignement), Han et al.('17) ; Carpentier and V.('17) (Sparsity Estimation) . . .

Towards a Minimax Lower Bound : Moment Matching

Two-point reduction to $\mathbb{P}_0$ and $\mathbf{P}_1 = \int \mathbb{P}_\theta \, d\mu_1(\theta)$ leads to suboptimal rate.

Need to consider composite-composite reduction of $\mathbf{P}_0 = \int \mathbb{P}_\theta \, d\mu_0(\theta)$ vs $\mathbf{P}_1$

Difficulty : $\chi^2(\mathbf{P}_0; \mathbf{P}_1) = \int \left(\frac{d\mathbf{P}_1}{d\mathbf{P}_0}\right)^2 d\mathbf{P}_0 - 1$ is not tractable

Workaround : Relating $\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}$ to measures of proximities between μ_0 and μ_1 :

Two (three) fruitful approaches :

- First moments of μ_0 and μ_1 are matching. (Lepski, Nemirovski and Spokoiny('99) for L_r norm).
  See also : Cai and Low('11) (L_1 norm) ; Jiao et al.('15) ; Wu and Yang('16) ; (Entropy Estimation) Bandeira et al.('17) (Multi-reference Alignement), Han et al.('17) ; Carpentier and V.('17) (Sparsity Estimation) . . .

Moment Matching

For $j = 0, 1$, take $\mu_j = \pi_j^{\otimes n}$

Lemma

$$\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}^2 \leq \frac{n}{\pi_1(\{0\})} \sum_{l \geq 1} \Big(\int x^l (d\pi_0(x) - d\pi_1(x)) \Big)^2 / l!$$

If

- Q first moments of π_0 and π_1 are equal
- $\mathrm{supp}(\pi_j) \subset [-M, M]$

then,

$$\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}^2 \leq \frac{2n}{\pi_1(\{0\})} \sum_{l > Q} \frac{M^{2l}}{l!}$$

Moment Matching

For $j = 0, 1$, take $\mu_j = \pi_j^{\otimes n}$

Lemma

$$\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}^2 \leq \frac{n}{\pi_1(\{0\})} \sum_{l \geq 1} \Big(\int x^l (d\pi_0(x) - d\pi_1(x)) \Big)^2 / l!$$

If

- Q first moments of π_0 and π_1 are equal
- $\mathrm{supp}(\pi_j) \subset [-M, M]$

then,

$$\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}^2 \leq \frac{2n}{\pi_1(\{0\})} \sum_{l > Q} \frac{M^{2l}}{l!}$$

$\rightsquigarrow \|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}$ is small for instance if $Q \geq \max(2eM^2, \log(n))$

Moment Matching : back to $m(\theta)$

Fix $\epsilon = k/(2n)$, $a_0 < a_1 = 0$,
$\pi_j = (1-\epsilon)\delta_{a_j} + \epsilon\nu_j$ and $\mathrm{supp}(\nu_\mathrm{j}) \subset [0, M]$.

$\rightsquigarrow$ if $\theta \sim \pi_j^{\otimes n}$, then w.h.p. $\theta \in \Theta_m[k]$ and $m(\theta) = a_j$.

Moment Matching : back to $m(\theta)$

Fix $\epsilon = k/(2n)$, $a_0 < a_1 = 0$,
$\pi_j = (1-\epsilon)\delta_{a_j} + \epsilon\nu_j$ and $\mathrm{supp}(\nu_{\mathrm{j}}) \subset [0, M]$.

$\rightsquigarrow$ if $\theta \sim \pi_j^{\otimes n}$, then w.h.p. $\theta \in \Theta_m[k]$ and $m(\theta) = a_j$.

Two points Method to $\int \mathbb{P}_\theta \, d\pi_j^{\otimes n}(\theta)$ $\Rightarrow$ Minimax Risk of $m(\theta)$ larger than $|a_0|$,
if enough moments of π_0 and π_1 are matching.

Moment Matching : back to $m(\theta)$

Fix $\epsilon = k/(2n)$, $a_0 < a_1 = 0$,
$\pi_j = (1-\epsilon)\delta_{a_j} + \epsilon\nu_j$ and $\mathrm{supp}(\nu_j) \subset [0, M]$.

$\rightsquigarrow$ if $\theta \sim \pi_j^{\otimes n}$, then w.h.p. $\theta \in \Theta_m[k]$ and $m(\theta) = a_j$.

Two points Method to $\int \mathbb{P}_\theta \, d\pi_j^{\otimes n}(\theta)$ $\Rightarrow$ Minimax Risk of $m(\theta)$ larger than $|a_0|$,
if enough moments of π_0 and π_1 are matching.

Desiderata

Find the largest $|a_0|$ such that two Probability measures ν_j supported on $[0, M]$ satisfy

$$\int_0^M x^r (d\nu_1 - d\nu_0) = \frac{\epsilon}{1-\epsilon} a_0^r, \qquad r = 1, \ldots, Q$$

Solution given by Hahn-Banach Theorem+ Riesz-Markov Theorem

Extremal Problem

Find the smallest $a_0 < 0$ such that

$$\sup_{P \in \mathcal{P}_Q : \|P\|_{\infty,[0,M]} \leq 1} |P(a_0) - P(0)| \leq \frac{2\epsilon}{1-\epsilon}$$

(Almost a Chebychev problem)

Conclusion : Lower bound

Theorem (Carpentier et al.('18))

For $k = 1, \dots, n$,

$$\inf_{\widehat{m}} \sup_{\theta \in \Theta_m[k]} \mathbb{E}_\theta \left|\widehat{m} - m(\theta)\right| \gtrsim \begin{cases} \frac{1}{\sqrt{n}} & \text{if } k \leq \sqrt{n} \ , \\ \frac{k}{n} \log^{-3/2}\left(1 + \frac{k}{\sqrt{n}}\right) & \text{if } \sqrt{n} < k \leq n/2 \ , \\ \frac{\log^2(n/(n-k))}{\log^{3/2}(n)} & \text{if } n/2 < k \leq n-1 \ , \end{cases}$$

- $k \leq \sqrt{n}$ $\rightsquigarrow$ Empirical Median is optimal
- $n - k \leq \sqrt{n}$ $\rightsquigarrow$ Empirical Min is optimal

Matching Upper bound : Moment approach

Two ideas : (similar to lower bounds)

- from Tests to Functional estimation
- Approximating $m(\theta)$.

Matching Upper bound : Moment approach

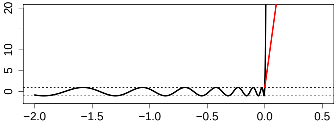
Two ideas : (similar to lower bounds)

- from Tests to Functional estimation
- Approximating $m(\theta)$.

Testing Problem : $m(\theta) = 0$ versus $m(\theta) < 0$

Population Analysis : Building a smooth function $\zeta(x)$ separating $\{x < 0\}$ from $\{x \geq 0\}$.

T_q : Chebychev Polynomial of degree q.
Take $\zeta(x) = T_q(2e^{\lambda x} - 1)$.



If $\theta_i \geq 0$, then $2e^{-\lambda\theta_i} - 1 \in [-1, 1]$ and $\zeta(-\theta_i) \in [-1, 1]$ for all i.
If $\theta_i < 0$, then $2e^{-\lambda\theta_i} - 1 > 1$ and $\zeta(-\theta_i) > 1$.

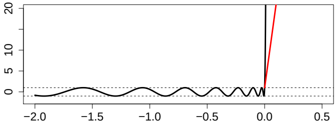
Two ideas : (similar to lower bounds)

- from Tests to Functional estimation
- Approximating $m(\theta)$.

Testing Problem : $m(\theta) = 0$ versus $m(\theta) < 0$

Population Analysis : Building a smooth function $\zeta(x)$ separating $\{x < 0\}$ from $\{x \geq 0\}$.

T_q : Chebychev Polynomial of degree q.
Take $\zeta(x) = T_q(2e^{\lambda x} - 1)$.



If $\theta_i \geq 0$, then $2e^{-\lambda\theta_i} - 1 \in [-1, 1]$ and $\zeta(-\theta_i) \in [-1, 1]$ for all i.
If $\theta_i < 0$, then $2e^{-\lambda\theta_i} - 1 > 1$ and $\zeta(-\theta_i) > 1$.

Unbiased Estimation of $\eta(\theta)$ + Multiple Tests $\rightsquigarrow$ Estimator $\widehat{m}_q$.

Theorem (Carpentier et al.('18))

For $k \in [\sqrt{n}, n - \sqrt{n}]$, take $q_k \asymp \log\left(\frac{k}{\sqrt{n}}\right)$, $\lambda_k = q_k^{-1/2}$. Then, $\widehat{m}_{q_k}$ is minimax.
Selection of $\widehat{m}_q$ (Lepski+Threshold) $\Rightarrow$ Minimax Adaptation

Extension : One-sided Contaminated Model

One-sided Contaminated Model

Y_i's are independent and either $Y_i \sim \mathcal{N}(m, 1)$ or $\mathcal{L}(Y_i) \overset{st.}{\gtrsim} \mathcal{N}(m, 1)$

$\rightsquigarrow$ k stands the number of "contaminated data".
$\rightsquigarrow$ one-sided counterpart of Huber's Contamination Model
(arises in multiple testing problems)

Extension : One-sided Contaminated Model

One-sided Contaminated Model

Y_i's are independent and either $Y_i \sim \mathcal{N}(m, 1)$ or $\mathcal{L}(Y_i) \overset{st.}{\gtrsim} \mathcal{N}(m, 1)$

$\rightsquigarrow$ k stands the number of "contaminated data".
$\rightsquigarrow$ one-sided counterpart of Huber's Contamination Model
(arises in multiple testing problems)

$\mathcal{M}_{k,\gtrsim}$: collection of such distributions.

Question : one-sided Contaminated Model vs Gaussian one-sided Model ?

Minimax Risk over the Contaminated Model

Theorem (Carpentier et al.('18))

For any $k = 1, \dots, n-1$,

$$\sup_{\widehat{m}} \inf_{\mathbf{P} \in \mathcal{M}_{k,\gtrsim}} \mathbf{E}\left[|\widehat{m} - m(\mathbf{P})|\right] \asymp \begin{cases} \frac{1}{\sqrt{n}} & \text{if } k \leq \sqrt{n} \ , \\ \frac{k}{n} \log^{-1/2}\left(1 + \frac{k}{\sqrt{n}}\right) & \text{if } \sqrt{n} < k \leq n/2 \ , \\ \frac{\log(n/(n-k))}{\log^{1/2}(n)} & \text{if } n/2 < k \leq n-1 \ , \end{cases}$$

Remark : At most a logarithmic difference with the Gaussian contaminated model

Remark : Matching upper bound achieved by suitable quantile estimators.
Lepski's method $\rightsquigarrow$ perfect Adaptation

Back to multiple Testing problems : Estimation of θ allows to correct the p-values and control the FDR as if θ_i was known in advance ($\log(n)^{-1/2}$ rate is required).

Take-Home Message

- Variations of two points methods for Adaptation.
- Moment Matching (and Fourier Matching) to handle composite-composite problems.

Take-Home Message

- Variations of two points methods for Adaptation.
- Moment Matching (and Fourier Matching) to handle composite-composite problems.

Thank you for your attention !