Minimax Estimation in Efron's Two-Groups Model

Chao Gao
University of Chicago

December 2023

Subhodh Kotekal

Stylized Story

z-scores

$$X_j \sim \begin{cases} N(0,1) & j \in \mathcal{H}_0 \\ N(\eta_j, 1) & j \in \mathcal{H}_1 \end{cases}$$

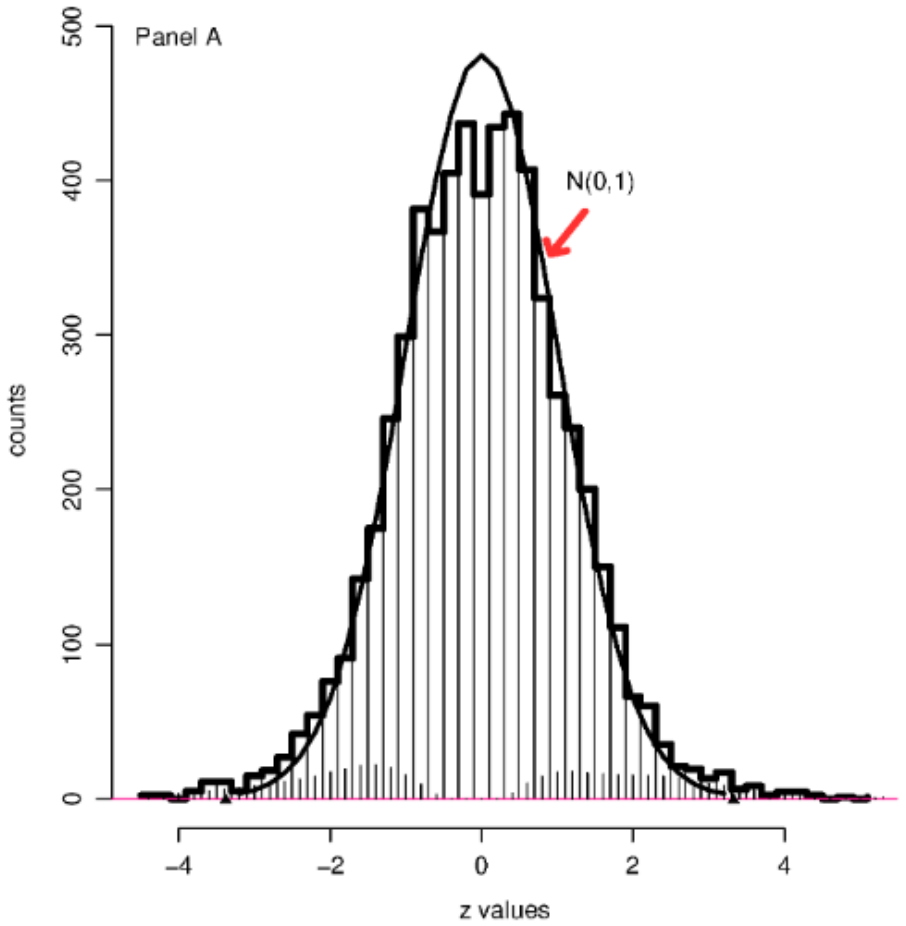


Figure 1: (Efron 2008) $n = 6033$ genes, prostate cancer study

Stylized story goes out of style

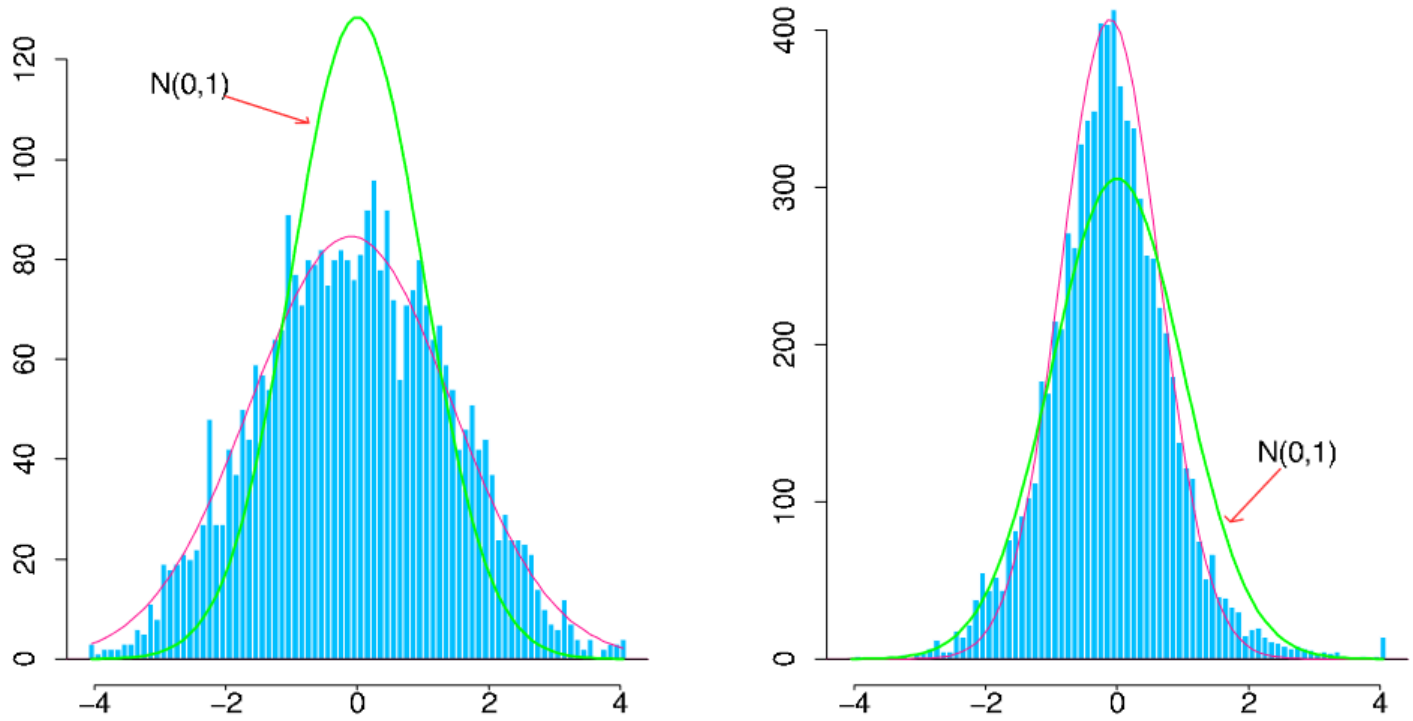


Figure 2: (Efron 2007) Left: $n = 3226$ genes, breast cancer study. Right: $n = 7680$ genes, HIV study.

$N(0, 1)$ appears inappropriate. $N(-0.09, 1.55^2)$ and $N(-0.11, 0.75^2)$ appear more appropriate.

Stylized story goes out of style

reasons for failure of theoretical null

correlation among z-scores,

unobserved covariates,

among others

Efron's suggested model

$$X_j \sim \begin{cases} N(\theta, \sigma^2) & j \in \mathcal{H}_0 \\ N(\eta_j, \sigma^2) & j \in \mathcal{H}_1 \end{cases}$$

[Efron 04]

Correlated z-Scores

one factor model

$$X_j = \eta_j + \sqrt{\rho}\,W + \sqrt{1-\rho}\,Z_j$$

$$\text{Cov}(X_i, X_j) = \begin{cases} 1 & i = j \\ \rho & i \neq j \end{cases}$$

$$\eta_j = 0 \text{ for all } j \in \mathcal{H}_0$$

marginally

$$X_j \sim \begin{cases} N(0,1) & j \in \mathcal{H}_0 \\ N(\eta_j, 1) & j \in \mathcal{H}_1 \end{cases}$$

conditionally

$$X_j | W \overset{\text{ind}}{\sim} \begin{cases} N(\sqrt{\rho}W, 1-\rho) & j \in \mathcal{H}_0 \\ N(\sqrt{\rho}W + \eta_j, 1-\rho) & j \in \mathcal{H}_1 \end{cases}$$

Robust Estimation

	Efron	**Huber**
frequentist	$X_j \sim \begin{cases} N(\theta, \sigma^2) & j \in \mathcal{H}_0 \\ N(\eta_j, \sigma^2) & j \in \mathcal{H}_1 \end{cases}$	$X_j \sim \begin{cases} N(\theta, \sigma^2) & j \in \mathcal{I} \\ \delta_{\eta_j} & j \in \mathcal{O} \end{cases}$
Bayesian	$(1-\epsilon)N(\theta, \sigma^2) + \epsilon N(0, \sigma^2) * Q$	$(1-\epsilon)N(\theta, \sigma^2) + \epsilon Q$

Robust Estimation

Theorem. Under Huber's setting

$$X_j \sim \begin{cases} N(\theta, \sigma^2) & j \in \mathcal{I} \\ \delta_{\eta_j} & j \in \mathcal{O} \end{cases}$$

or

$$X_j \sim (1-\epsilon)N(\theta, \sigma^2) + \epsilon Q$$

[Chen, Gao & Ren 15]

Robust Estimation

Efron

$$\frac{9}{10}N(\theta, \sigma^2) + \frac{1}{10}N(0, \sigma^2) * Q$$

identifiable

Huber

$$\frac{9}{10}N(\theta, \sigma^2) + \frac{1}{10}Q$$

not identifiable

The null distribution can be consistently estimated under Efron's setting, even when the non-null proportion is constant.

Goal

An alternative way of writing Efron's model

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2)$$

sparse non-null effects

$$\|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

Question:

What is the minimax rate of estimation of the null?

Literature

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

penalized estimation

$$\min_{\theta} \min_{\gamma \in \mathbb{R}^n} \left[\frac{1}{2} \sum_{j=1}^{n} (X_j - \theta - \gamma_j)^2 + \rho \|\gamma\|_1 \right]$$

equivalent form

$$\min_{\theta} \sum_{j=1}^{n} \mathcal{L}_{\text{Huber}}(X_j - \theta; \rho)$$

$$\mathcal{L}_{\text{Huber}}(t; \rho) = \begin{cases} \dfrac{t^2}{2} & |t| \leq \rho \\ \rho|t| - \dfrac{\rho^2}{2} & |t| > \rho \end{cases}$$

[Gannaz 07, Antoniadis 07, She & Owen 11]

Literature

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

penalized estimation

$$\min_{\theta} \min_{\gamma \in \mathbb{R}^n} \left[\frac{1}{2} \sum_{j=1}^{n} (X_j - \theta - \gamma_j)^2 + \rho \|\gamma\|_1 \right]$$

Collier & Dalalyan (2019)

$$|\widehat{\theta} - \theta|^2 = O_P\left(\frac{1}{n} + \frac{k^2}{n^2} \log n\right)$$

Literature

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

characteristic function

$$\frac{1}{n} \sum_{j=1}^{n} e^{itX_j}$$

Literature

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

characteristic function

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\right) = e^{-\frac{t^2\sigma^2}{2}}\left[\left(1-\frac{k}{n}\right)e^{it\theta} + \frac{k}{n}\widehat{Q}(t)\right]$$

Cai & Jin (2010) $$|\widehat{\theta} - \theta|^2 = O_P\left(\frac{1}{n} + \frac{k^2}{n^2}\frac{1}{(\log n)^{\alpha+1}}\right)$$

Main Result

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

Theorem [Kotekal & G.]. The minimax rate of mean estimation is

$$\epsilon(k,n)^2 \asymp \begin{cases} \dfrac{1}{n} & 1 \leq k < \sqrt{n} \\ \dfrac{k^2}{n^2} \dfrac{1}{\log\left(1 + \frac{k^2}{n}\right)} & \sqrt{n} \leq k < \dfrac{n}{4} \\ \dfrac{1}{\log\left(1 + \frac{(n-2k)^2}{n}\right)} & \dfrac{n}{4} \leq k < \dfrac{n}{2} - \sqrt{n} \\ \log\left(1 + \dfrac{n}{(n-2k)^2}\right) & \dfrac{n}{2} - \sqrt{n} \leq k < \dfrac{n}{2} \end{cases}$$

[Carpentier & Verzelen]

Main Result

$$\epsilon(k,n)^2 \asymp \begin{cases} \dfrac{1}{n} & 1 \le k < \sqrt{n} \\ \dfrac{k^2}{n^2}\dfrac{1}{\log\left(1+\frac{k^2}{n}\right)} & \sqrt{n} \le k < \dfrac{n}{4} \\ \dfrac{1}{\log\left(1+\frac{(n-2k)^2}{n}\right)} & \dfrac{n}{4} \le k < \dfrac{n}{2} - \sqrt{n} \\ \log\left(1+\dfrac{n}{(n-2k)^2}\right) & \dfrac{n}{2} - \sqrt{n} \le k < \dfrac{n}{2} \end{cases}$$

Conclusion 1

parametric rate $\Longleftrightarrow$ $|\mathcal{H}_1| \lesssim \sqrt{n}$

Main Result

$$\epsilon(k,n)^2 \asymp \begin{cases} \dfrac{1}{n} & 1 \le k < \sqrt{n} \\ \dfrac{k^2}{n^2}\dfrac{1}{\log\left(1+\frac{k^2}{n}\right)} & \sqrt{n} \le k < \dfrac{n}{4} \\ \dfrac{1}{\log\left(1+\frac{(n-2k)^2}{n}\right)} & \dfrac{n}{4} \le k < \dfrac{n}{2} - \sqrt{n} \\ \log\left(1+\dfrac{n}{(n-2k)^2}\right) & \dfrac{n}{2} - \sqrt{n} \le k < \dfrac{n}{2} \end{cases}$$

Conclusion 2

logarithmic rate $|\widehat{\theta} - \theta|^2 \lesssim \dfrac{1}{\log n}$ when $\dfrac{|\mathcal{H}_1|}{n} \in \left\{\dfrac{1}{100}, \dfrac{1}{10}, \dfrac{1}{3}, 0.49\right\}$

Main Result

$$\epsilon(k,n)^2 \asymp \begin{cases} \dfrac{1}{n} & 1 \le k < \sqrt{n} \\ \dfrac{k^2}{n^2}\dfrac{1}{\log\left(1+\frac{k^2}{n}\right)} & \sqrt{n} \le k < \dfrac{n}{4} \\ \dfrac{1}{\log\left(1+\frac{(n-2k)^2}{n}\right)} & \dfrac{n}{4} \le k < \dfrac{n}{2} - \sqrt{n} \\ \log\left(1+\dfrac{n}{(n-2k)^2}\right) & \dfrac{n}{2} - \sqrt{n} \le k < \dfrac{n}{2} \end{cases}$$

Conclusion 3

consistency $\Longleftrightarrow$ $\dfrac{n-2k}{\sqrt{n}} = \dfrac{|\mathcal{H}_0| - |\mathcal{H}|_1}{\sqrt{n}} \to \infty$

Main Result

$$\epsilon(k,n)^2 \asymp \begin{cases} \dfrac{1}{n} & 1 \le k < \sqrt{n} \\ \dfrac{k^2}{n^2}\dfrac{1}{\log\left(1+\frac{k^2}{n}\right)} & \sqrt{n} \le k < \dfrac{n}{4} \\ \dfrac{1}{\log\left(1+\frac{(n-2k)^2}{n}\right)} & \dfrac{n}{4} \le k < \dfrac{n}{2} - \sqrt{n} \\ \log\left(1+\dfrac{n}{(n-2k)^2}\right) & \dfrac{n}{2} - \sqrt{n} \le k < \dfrac{n}{2} \end{cases}$$

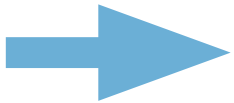
Conclusion 4

diverging rate $|\widehat{\theta} - \theta|^2 \lesssim \log n$ when $\dfrac{n}{2} - n^{0.49} < |\mathcal{H}_1| < \dfrac{n}{2}$

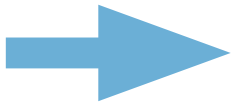
Characteristic Function

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\right) = e^{it\theta - \frac{t^2}{2}}\left(1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}\right)$$



$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\theta)+\frac{t^2}{2}}\right) = 1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}$$

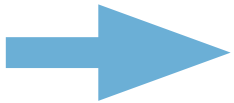


$$\theta \stackrel{?}{=} \operatorname*{argmin}_{\mu} \left|\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\mu)+\frac{t^2}{2}}\right) - \left(1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}\right)\right|$$

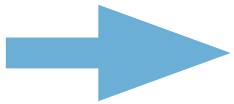
Characteristic Function

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\right) = e^{it\theta - \frac{t^2}{2}}\left(1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}\right)$$



$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\theta)+\frac{t^2}{2}}\right) = 1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}$$



$$\theta = \underset{\mu}{\operatorname{argmin}} \sup_{t\in\mathbb{R}} \left|\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\mu)+\frac{t^2}{2}}\right) - \left(1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}\right)\right|$$

Characteristic Function

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\right) = e^{it\theta - \frac{t^2}{2}}\left(1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}\right)$$

$$\Rightarrow \quad \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\theta)+\frac{t^2}{2}}\right) = 1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}$$

$$\Rightarrow \quad \theta = \operatorname*{argmin}_{\mu} \sup_{t\in\mathbb{R}} \left|\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\mu)+\frac{t^2}{2}}\right) - \left(1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}\right)\right|$$

$$\theta = \operatorname*{argmin}_{\mu} \sup_{t\in\mathbb{R}} \inf_{\substack{\zeta\in\mathbb{C} \\ |\zeta|\leq 1}} \left|\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\mu)+\frac{t^2}{2}}\right) - \left(1 - \frac{k}{n} + \frac{k}{n}\zeta\right)\right|$$

Characteristic Function

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\right) = e^{it\theta - \frac{t^2}{2}}\left(1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}\right)$$

$$\Rightarrow \quad \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\theta)+\frac{t^2}{2}}\right) = 1 - \frac{k}{n} + \frac{k}{n}\frac{1}{k}\sum_{j\in\mathcal{H}_1} e^{it\gamma_j}$$

Proposition. When $\frac{k}{n} < \frac{1}{2}$, the following holds for arbitrary non-null effects,

$$\theta = \operatorname*{argmin}_{\mu} \sup_{t\in\mathbb{R}} \inf_{\substack{\zeta\in\mathbb{C} \\ |\zeta|\leq 1}} \left| \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it(X_j-\mu)+\frac{t^2}{2}}\right) - \left(1 - \frac{k}{n} + \frac{k}{n}\zeta\right)\right|$$

Characteristic Function

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\theta = \underset{\mu}{\operatorname{argmin}} \sup_{t \in \mathbb{R}} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^{n} e^{it(X_j - \mu) + \frac{t^2}{2}} \right) - \left(1 - \frac{k}{n} + \frac{k}{n}\zeta \right) \right|$$

$$\widehat{\theta} = \underset{\mu}{\operatorname{argmin}} \sup_{|t| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{it(X_j - \mu) + \frac{t^2}{2}} - \left(1 - \frac{k}{n} + \frac{k}{n}\zeta \right) \right|$$

bias $\frac{k}{n\tau}$

Characteristic Function

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\theta = \operatorname*{argmin}_{\mu} \sup_{t \in \mathbb{R}} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^{n} e^{it(X_j - \mu) + \frac{t^2}{2}} \right) - \left(1 - \frac{k}{n} + \frac{k}{n}\zeta \right) \right|$$

$$\widehat{\theta} = \operatorname*{argmin}_{\mu} \sup_{|t| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{it(X_j - \mu) + \frac{t^2}{2}} - \left(1 - \frac{k}{n} + \frac{k}{n}\zeta \right) \right|$$

stochastic error $\frac{1}{\sqrt{n}} e^{\tau^2/2}$

Characteristic Function

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\widehat{\theta} = \underset{\mu}{\operatorname{argmin}} \sup_{|t| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{it(X_j - \mu) + \frac{t^2}{2}} - \left(1 - \frac{k}{n} + \frac{k}{n}\zeta\right) \right|$$

bias $\frac{k}{n\tau}$ stochastic error $\frac{1}{\sqrt{n}} e^{\tau^2/2}$

set $\tau \asymp 1 + \sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}$ → achieves optimality when $1 \leq k < \frac{n}{2} - \sqrt{n}$

Kernel Mode Estimator

$$X_j = \theta + \gamma_j + Z_j \sim N(\theta + \gamma_j, 1) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

[Parzen 62, Eddy 80 & 82]
[Efron 04]

$$\widehat{\theta} \;=\; \underset{\mu}{\operatorname{argmax}} \sum_{j=1}^{n} \mathbb{I}\{|X_j - \mu| \leq h\}$$

a widening bandwidth

Proposition. Suppose $\frac{n}{2} - \sqrt{n} \leq k < \frac{n}{2}$. Take $h \asymp \sqrt{\log\left(1 + \frac{n}{(n-2k)^2}\right)}$. Then, w.h.p.,

$$|\widehat{\theta} - \theta|^2 \lesssim \log\left(1 + \frac{n}{(n-2k)^2}\right)$$

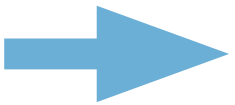
Lower Bound

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(Q_0 * N(0,1)) = p_0$$

$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(\theta,1) + \varepsilon(Q_1 * N(0,1)) = p_1$$

Goal: find θ for which there exist Q_0 and Q_1 such that $\chi^2(p_0 \| p_1) \lesssim n^{-1}$

a simple observation

when $\epsilon = \frac{1}{2}$, take $\begin{cases} Q_0 = \delta_\theta \\ Q_1 = \delta_0 \end{cases}$  $\chi^2(p_0 \| p_1) = 0$

Lower Bound

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(Q_0 * N(0,1)) = p_0$$

$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(\theta,1) + \varepsilon(Q_1 * N(0,1)) = p_1$$

Goal: find θ for which there exist Q_0 and Q_1 such that $\chi^2(p_0 \| p_1) \lesssim n^{-1}$

Fourier transform: $\chi^2(p_0 \| p_1) \leq \frac{\sqrt{2\pi}}{1-\epsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int |\hat{p}_1^{(k)}(t) - \hat{p}_0^{(k)}(t)|^2 dt$

$$\begin{cases} \hat{p}_0(t) = e^{-t^2/2}(1 - \epsilon + \epsilon \hat{q}_0(t)) \\ \hat{p}_1(t) = e^{-t^2/2}((1-\epsilon)e^{-i\theta t} + \epsilon \hat{q}_1(t)) \end{cases}$$

Lower Bound

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(Q_0 * N(0,1)) = p_0$$

$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(\theta,1) + \varepsilon(Q_1 * N(0,1)) = p_1$$

Goal: find θ for which there exist Q_0 and Q_1 such that $\chi^2(p_0 \| p_1) \lesssim n^{-1}$

Fourier transform: $\chi^2(p_0 \| p_1) \leq \frac{\sqrt{2\pi}}{1-\epsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int |\hat{p}_1^{(k)}(t) - \hat{p}_0^{(k)}(t)|^2 dt$

hope $\hat{q}_1(t) - \hat{q}_0(t) = \frac{1-\epsilon}{\epsilon}(1 - e^{-i\theta t})$

this is impossible

Lower Bound

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(Q_0 * N(0,1)) = p_0$$

$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(\theta,1) + \varepsilon(Q_1 * N(0,1)) = p_1$$

Goal: find θ for which there exist Q_0 and Q_1 such that $\chi^2(p_0 \| p_1) \lesssim n^{-1}$

Fourier transform: $\chi^2(p_0 \| p_1) \leq \frac{\sqrt{2\pi}}{1-\epsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int |\hat{p}_1^{(k)}(t) - \hat{p}_0^{(k)}(t)|^2 dt$

Proposition. Suppose $\theta \asymp \epsilon/\tau$. There exist

$$\tilde{q}_1(t) - \tilde{q}_0(t) = \begin{cases} \frac{1-\epsilon}{\epsilon}(1 - e^{-i\theta t}) & |t| \leq \tau \\ \dots & \tau < |t| \leq 2\tau \\ 0 & |t| > 2\tau \end{cases}$$

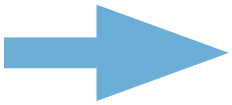
Lower Bound

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(Q_0 * N(0,1)) = p_0$$

$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(\theta,1) + \varepsilon(Q_1 * N(0,1)) = p_1$$

Goal: find θ for which there exist Q_0 and Q_1 such that $\chi^2(p_0 \| p_1) \lesssim n^{-1}$

Fourier transform: $\chi^2(p_0 \| p_1) \leq \frac{\sqrt{2\pi}}{1-\epsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int |\hat{p}_1^{(k)}(t) - \hat{p}_0^{(k)}(t)|^2 dt$



$\chi^2(p_0 \| p_1) \lesssim \epsilon^2 e^{-\tau^2/4}$

but $\chi^2(p_0 \| p_1) \neq 0$ even when $\epsilon = \frac{1}{2}$

Lower Bound

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(Q_0 * N(0,1)) = p_0$$

$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(\theta,1) + \varepsilon(Q_1 * N(0,1)) = p_1$$

Goal: find θ for which there exist Q_0 and Q_1 such that $\chi^2(p_0 \| p_1) \lesssim n^{-1}$

Fourier transform: $\chi^2(p_0 \| p_1) \leq \frac{\sqrt{2\pi}}{1-\epsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int |\hat{p}_1^{(k)}(t) - \hat{p}_0^{(k)}(t)|^2 dt$

Proposition. Suppose $\theta \asymp \epsilon/\tau$. There exist

$$\hat{q}_1(t) - \hat{q}_0(t) = \begin{cases} \frac{1-\epsilon}{\epsilon}(1 - e^{-i\theta t}) & |t| \leq \tau \\ \dots & \tau < |t| \leq 2\tau \\ 2\epsilon(1 - e^{-i\theta t}) & |t| > 2\tau \end{cases}$$

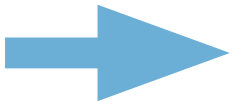
Lower Bound

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(0,1) + \varepsilon(Q_0 * N(0,1)) = p_0$$

$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1-\varepsilon)N(\theta,1) + \varepsilon(Q_1 * N(0,1)) = p_1$$

Goal: find θ for which there exist Q_0 and Q_1 such that $\chi^2(p_0 \| p_1) \lesssim n^{-1}$

Fourier transform: $\chi^2(p_0 \| p_1) \leq \frac{\sqrt{2\pi}}{1-\epsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int |\hat{p}_1^{(k)}(t) - \hat{p}_0^{(k)}(t)|^2 dt$



$$\chi^2(p_0 \| p_1) \lesssim (1-2\epsilon)^2 \epsilon^2 e^{-\tau^2/4}$$

take $\tau^2 \asymp \log\left(1 + n\epsilon^2(1-2\epsilon)^2\right) = \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)$

Estimating Variance

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\widehat{N}(t) = \left| \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} \right| \qquad N(t) = \left| e^{it\theta - \frac{t^2\sigma^2}{2}} \left(1 - \frac{k}{n} + \frac{k}{n} \frac{1}{k} \sum_{j \in \mathcal{H}_1} e^{it\gamma_j} \right) \right|$$

Estimating Variance

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\widehat{N}(t) = \left| \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} \right| \qquad -\frac{2 \log N(t)}{t^2} = \sigma^2 \boxed{-\frac{2}{t^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{H}_1} e^{it\gamma_j} \right) \right|}$$

positive bias

$$\widehat{\sigma}^2 = -\frac{2 \log \widehat{N}(t)}{t^2}$$

[Comminges, Collier, Ndaoud & Tsybakov 21]

Estimating Variance

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\widehat{N}(t) = \left| \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} \right| \qquad -\frac{2 \log N(t)}{t^2} = \sigma^2 \boxed{-\frac{2}{t^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{H}_1} e^{it\gamma_j} \right) \right|}$$

positive bias

$$\left| 1 - \frac{1}{k} \sum_{j \in \mathcal{H}_1} e^{it\gamma_j} \right|^2$$

Proposition. For any $\tau > 0$

$$\sup_{t \in [\tau, 10\tau]} \frac{1}{k} \sum_{j \in \mathcal{H}_1} \cos(t\gamma_j) \geq -\frac{1}{5}$$

Estimating Variance

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\widehat{N}(t) = \left| \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} \right| \qquad -\frac{2 \log N(t)}{t^2} = \sigma^2 \left| -\frac{2}{t^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{H}_1} e^{it\gamma_j} \right) \right| \right|$$

positive bias

$$\widehat{\sigma}^2 = \inf_{t \in [\tau, 10\tau]} -\frac{2 \log \widehat{N}(t)}{t^2}$$

Proposition. For any $\tau > 0$

$$\sup_{t \in [\tau, 10\tau]} \frac{1}{k} \sum_{j \in \mathcal{H}_1} \cos(t\gamma_j) \geq -\frac{1}{5}$$

Estimating Variance

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

$$\widehat{N}(t) = \left| \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} \right| \qquad \widehat{\sigma}^2 = \inf_{t \in [\tau, 10\tau]} -\frac{2 \log \widehat{N}(t)}{t^2}$$

Theorem [Kotekal & G.]. Set $\tau^2 \asymp \frac{1}{\sigma^2} \log\left(1 + \frac{k^2}{n}\right)$. The estimator achieves the following minimax rate,

$$\frac{|\widehat{\sigma}^2 - \sigma^2|^2}{\sigma^4} \lesssim \begin{cases} \frac{1}{n} & 1 \leq k < \sqrt{n} \\ \frac{k^2}{n^2} \frac{1}{\left(\log\left(1 + \frac{k^2}{n}\right)\right)^2} & \sqrt{n} < k < \frac{n}{2} \end{cases}$$

Estimating Null Distribution

$$X_j = \theta + \gamma_j + \sigma Z_j \sim N(\theta + \gamma_j, \sigma^2) \qquad \|\gamma\|_0 = \sum_{j=1}^{n} \mathbb{I}\{\gamma_j \neq 0\} \leq k$$

Theorem [Kotekal & G.]. With both location and variance estimators, $N(\widehat{\theta}, \widehat{\sigma}^2)$ achieves the following minimax rate

$$\mathsf{TV}\left(N(\widehat{\theta}, \widehat{\sigma}^2), N(\theta, \sigma^2)\right)^2 \lesssim \begin{cases} \frac{1}{n} & 1 \leq k < \sqrt{n} \\ \frac{k^2}{n^2} \frac{1}{\log\left(1 + \frac{k^2}{n}\right)} & \sqrt{n} \leq k < \frac{n}{4} \\ \frac{1}{\log\left(1 + \frac{(n-2k)^2}{n}\right)} & \frac{n}{4} \leq k < \frac{n}{2} - \sqrt{n} \\ 1 & k \geq \frac{n}{2} - \sqrt{n} \end{cases}$$

Thank You