Butlletí de la Societat Catalana de Matemàtiques
Vol. 37, núm. 1, 2022. Pàg. 61–93. DOI: 10.2436/20.2002.01.103

Estimació de la mitjana: una revisió d'avenços recents

Gábor Lugosi

Resum: Segurament el problema estadístic més bàsic és estimar el valor esperat d'una variable aleatòria basant-ho en observacions independents de la mateixa distribució. Aquesta pregunta aparentment senzilla planteja dificultats sorprenents tant des del punt de vista estadístic com computacional. En aquest article analitzem alguns dels avenços recents en l'estimació de la mitjana. En particular, es descriuen estimadors de mitjanes subgaussians per a dades que poden tenir cues pesades, tant en configuracions univariants com multivariants. Ens centrem en estimadors basats en tècniques de medianes de mitjanes però també es revisen altres mètodes com la mitjana retallada i l'estimador de Catoni. Presentem demostracions detallades dels resultats bàsics.

Paraules clau: estimació de la mitjana, distribucions de cua pesada, robustesa, aprenentatge estadístic.

Classificació MSC2010: 62G05, 62G15, 62G35.

1 Introducció

Sens dubte, el problema més fonamental de l'estadística és estimar el valor esperat μ d'una variable aleatòria X basant-se en una mostra de n observacions independents i idènticament distribuïdes extretes d'una distribució de X. L'elecció òbvia d'estimador és, per descomptat, la mitjana empírica. Les seves propietats són ben conegudes a partir de resultats clàssics de la teoria de la probabilitat. Tanmateix, des dels inicis, els experts en estadística s'han preocupat per la qualitat de la mitjana empírica, especialment quan la distribució pot ser de *cua pesada* o hi pot haver *dades atípiques*. Aquesta preocupació va donar lloc a l'àrea de l'*estadística robusta*, que aborda el problema d'estimació de la mitjana (i altres problemes estadístics) per a aquest tipus de dades. Les referències clàssiques inclouen Huber [31], Huber i Ronchetti [32], Hampel, Ronchetti, Rousseeuw i Stahel [23] i Tukey [59].

Motivat per aplicacions en aprenentatge automàtic i en ciència de dades, en els darrers anys hi ha hagut un interès creixent en la construcció d'estimacions

de la mitjana i de funcions de regressió amb el requisit que els estimadors hagin d'aconseguir *alta precisió* amb una gran *confiança*. Avui dia, s'entén molt millor quin és el millor balanç precisió/confiança que es pot assolir i l'objectiu d'aquest article és analitzar-ne alguns dels avenços recents. Ens centrarem en el problema d'estimació de la mitjana, tant en configuracions univariants com multivariants. Presentem una discussió detallada sobre quin és el millor rendiment que hom pot esperar, sobre com descriure una varietat d'estimadors i sobre com analitzar-ne el rendiment. Parem una atenció especial a una metodologia senzilla però potent basada en tècniques de *medianes de mitjanes*.

A la secció 2, abordem el problema més simple, el d'estimació de la mitjana univariant. Ens centrem en estimadors *subgaussians* i n'explorem les possibilitats i limitacions. La secció 3 es dedica al problema multivariant, que és substancialment més difícil. Ampliem la noció d'estimador subgaussià en el context multivariant i analitzem diversos estimadors.

Per a una revisió més detallada, remetem el lector al recent article de Lugosi i Mendelson [45].

2 Estimació de la mitjana d'una variable aleatòria real

Comencem examinant el problema clàssic d'estimar la mitjana d'una variable aleatòria. Siguin $X_1, \dots, X_n$ variables aleatòries independents, distribuïdes de manera idèntica, amb mitjana $\mu = \mathbb{E}X_1$. Es pretén estimar μ a partir d'observacions d'aquestes variables aleatòries. Un estimador $\hat{\mu}_n = \hat{\mu}_n(X_1, \dots, X_n)$ és simplement una funció (mesurable) de $X_1, \dots, X_n$.

Per descomptat, l'opció més natural d'un estimador de la mitjana és la mitjana empírica estàndard

$$\overline{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Tanmateix, tal com argumentem a continuació, hi pot haver altres estimadors més adients, segons quin sigui l'objectiu.

La qualitat d'un estimador es pot mesurar de diverses maneres. Els primers treballs estadístics es van centrar, en la seva majoria, en mesures del risc esperat, com ara l'*error quadràtic mitjà*

$$\mathbb{E}[(\hat{\mu}_n - \mu)^2].$$

Si hom se centra en l'error quadràtic mitjà, aleshores la mitjana empírica té un rendiment excel·lent. Suposem que els X_i tenen un segon moment finit i $\sigma^2 = \mathbb{E}[(X_i - \mu)^2]$ denota la seva variància. Aleshores, és un exercici fàcil veure que l'error quadràtic mitjà de $\overline{\mu}_n$ és igual a σ^2/n. Al mateix temps, això és òptim fins i tot si suposem que la distribució de X_i és gaussiana. De fet, del teorema clàssic de Rao-Blackwell ([55, 4, 36]) es desprèn que no es pot esperar obtenir un error quadràtic mitjà més petit quan s'estima la mitjana d'una variable aleatòria gaussiana.

No obstant això, les mesures de risc com l'error quadràtic mitjà poden induir a error. De fet, si la diferència $|\widehat{\mu}_n - \mu|$ no és prou concentrada, aleshores el valor esperat no necessàriament reflecteix el comportament «típic» de l'error. Per motius com aquest, sovint preferim els estimadors $\widehat{\mu}_n$, que són propers a μ *amb alta probabilitat*. Així, el nostre objectiu és entendre, per a qualsevol mida de mostra n i el paràmetre de confiança $\delta \in (0,1)$, el valor més petit possible $\epsilon = \epsilon(n,\delta)$ tal que

$$\mathbb{P}\{|\widehat{\mu}_n - \mu| > \epsilon\} \leq \delta. \tag{1}$$

És important destacar que (1) és un criteri no asimptòtic: hom voldria obtenir estimacions quantitatives sobre com la precisió ϵ s'escala amb el paràmetre de confiança δ i la mida de la mostra n. Aquest tipus d'estimació recorda el *PAC* (acrònim de *probablement aproximadament correcte*), un marc adoptat habitualment en la teoria de l'aprenentatge estadístic; vegeu Valiant [61], Vapnik i Chervonenkis [63] i Blumer, Ehrenfeucht, Haussler i Warmuth [5].

Les propietats de la mitjana empírica s'entenen molt bé, ja que les mitjanes de variables aleatòries independents s'han estudiat des dels inicis de la teoria de la probabilitat. Una pedra angular d'aquesta teoria és el teorema del límit central. Aquest teorema garanteix que la mitjana empírica té cues gaussianes, asimptòticament, quan n tendeix a infinit. Més precisament,

$$\mathbb{P}\left\{|\overline{\mu}_n - \mu| > \frac{\sigma\Phi^{-1}(1-\delta/2)}{\sqrt{n}}\right\} \to \delta,$$

on $\Phi(x) = \mathbb{P}\{G \leq x\}$ és la funció de distribució acumulada d'una variable aleatòria normal estàndard G. Es pot veure fàcilment (per exemple, utilitzant el fet que per a $t \geq 1$, $\exp(-t^2/2) \leq t\exp(-t^2/2)$), que per a tots els $x \geq 0$,

$$1 - \Phi(x) \leq e^{-x^2/2}.$$

Això implica que $\Phi^{-1}(1-\delta/2) \leq \sqrt{2\log(2/\delta)}$, i el teorema del límit central afirma que

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{\mu}_n - \mu| > \frac{\sigma\sqrt{2\log(2/\delta)}}{\sqrt{n}}\right\} \leq \delta.$$

No obstant això, aquesta és una estimació asimptòtica i no la quantitativa que esperàvem. Tot i així, el nostre objectiu és obtenir fites de rendiment no asimptòtiques de la mateixa forma. En particular, diem que un estimador de la mitjana $\widehat{\mu}_n$ és *L-subgaussià* si existeix una constant $L > 0$ de manera que per a totes les mides de mostra n i amb probabilitat almenys $1-\delta$, se satisfà

$$|\widehat{\mu}_n - \mu| \leq \frac{L\sigma\sqrt{\log(2/\delta)}}{\sqrt{n}}. \tag{2}$$

Val la pena esmentar, en aquest punt, el fet conegut que, si tot el que se sap és que la distribució desconeguda és gaussiana, llavors la mitjana empírica

és òptima per a totes les mides de mostra n i tots els nivells de confiança δ (vegeu Catoni [8, proposició 6.1] per a un enunciat precís). A més, la següent observació, establerta per Devroye, Lerasle, Lugosi i Oliveira [17], mostra que (2) és essencialment el millor que es pot esperar en general, fins i tot si hom està interessat en un nivell de confiança fixat:

TEOREMA 1. *Sigui $n > 5$ un número enter positiu. Siguin $\mu \in \mathbb{R}$, $\sigma > 0$ i $\delta \in (2e^{-n/4}, 1/2)$. Aleshores, per a qualsevol estimador de la mitjana $\hat{\mu}_n$, existeix una distribució amb mitjana μ i variància σ^2 tal que*

$$\mathbb{P}\left\{|\hat{\mu}_n - \mu| > \sigma\sqrt{\frac{\log(1/\delta)}{n}}\right\} \geq \delta.$$

PROVA. Per trobar la fita inferior, n'hi ha prou de considerar dues distribucions, P_+, P_-, ambdues concentrades en dos punts, definits per

$$P_+(\{0\}) = P_-(\{0\}) = 1 - p, \quad P_+(\{c\}) = P_-(\{-c\}) = p,$$

on $p \in [0,1]$ i $c > 0$. Observem que les mitjanes de les dues distribucions són $\mu_{P_+} = pc$ i $\mu_{P_-} = -pc$ i que totes dues tenen variància $\sigma^2 = c^2 p(1-p)$.

Per a $i = 1, \ldots, n$, considerem parells independents de variables aleatòries de valors reals (X_i, Y_i) tals que

$$\mathbb{P}\{X_i = Y_i = 0\} = 1 - p \quad \text{i} \quad \mathbb{P}\{X_i = c, Y_i = -c\} = p.$$

Fixem-nos que X_i es distribueix com P_+ i Y_i es distribueix com P_-. Sigui $\delta \in (0, 1/2)$. Si $\delta \geq 2e^{-n/4}$ i $p = (1/(2n))\log(2/\delta)$, aleshores (utilitzant $1 - p \geq \exp(-p/(1-p))$),

$$\mathbb{P}\{(X_1, \ldots, X_n) = (Y_1, \ldots, Y_n)\} = (1-p)^n \geq 2\delta.$$

Sigui $\hat{\mu}_n$ un estimador de la mitjana qualsevol, que possiblement depèn de δ. Aleshores,

$$\begin{aligned}
&\max(\mathbb{P}\{|\hat{\mu}_n(X_1^n) - \mu_{P_+}| > cp\}, \mathbb{P}\{|\hat{\mu}_n(Y_1^n) - \mu_{P_-}| > cp\}) \geq \\
&\quad \geq \frac{1}{2}\mathbb{P}\{|\hat{\mu}_n(X_1, \ldots, X_n) - \mu_{P_+}| > cp \quad \text{o} \quad |\hat{\mu}_n(Y_1, \ldots, Y_n) - \mu_{P_-}| > cp\} \geq \\
&\quad \geq \frac{1}{2}\mathbb{P}\{\hat{\mu}_n(X_1, \ldots, X_n) = \hat{\mu}_n(Y_1, \ldots, Y_n)\} \geq \\
&\quad \geq \frac{1}{2}\mathbb{P}\{(X_1, \ldots, X_n) = (Y_1, \ldots, Y_n)\} \geq \delta.
\end{aligned}$$

A partir de $\sigma^2 = c^2 p(1-p)$ i $p \leq 1/2$, tenim que $cp \geq \sigma\sqrt{p/2}$, i per tant,

$$\max\left(\mathbb{P}\left\{|\hat{\mu}_n(X_1,\ldots,X_n) - \mu_{P_+}| > \sigma\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right\},\right.$$
$$\left.\mathbb{P}\left\{|\hat{\mu}_n(Y_1,\ldots,Y_n) - \mu_{P_-}| > \sigma\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right\}\right) \geq \delta.$$

D'aquí se'n desprèn el teorema 1. □

Considerant el teorema 1, el nostre objectiu és dissenyar estimadors que tinguin un rendiment amb una taxa d'error subgaussiana.

Naturalment, la primera tasca a fer és comprovar si l'elecció més òbvia d'estimador de la mitjana —la mitjana empírica— és L-subgaussià per a algun L. D'una banda, és fàcil veure que, sota determinades condicions sobre la distribució de X_i, la mitjana empírica mostra un rendiment subgaussià. De fet, si els X_i són tals que existeix $L > 0$ tal que per a tot $\lambda > 0$

$$\mathbb{E}e^{\lambda(X_i-\mu)} \leq e^{\sigma^2\lambda^2/L^2},$$

aleshores la mitjana empírica $\overline{\mu}_n$ és L-subgaussiana per a tota $\delta \in (0,1)$, tal com es veu fàcilment amb la fita de Chernoff.

D'altra banda, els supòsits d'aquest tipus són força restrictius i imposen condicions fortes sobre el decaïment de la probabilitat de la cua de les X_i. Quan les X_i no presenten aquest decaïment de la cua, la mitjana empírica no té per què ser subgaussiana.

Per exemple, si assumim només que σ existeix (és a dir, que la variància de X_i és finita), llavors es pot utilitzar la desigualtat de Txebixev. Aquesta desigualtat implica que, amb probabilitat almenys $1-\delta$,

$$|\overline{\mu}_n - \mu| \leq \sigma\sqrt{\frac{1}{n\delta}}. \tag{3}$$

No obstant això, sota el supòsit de la finitud del segon moment, això és essencialment el millor que es pot esperar. Tot i que la fita de (3) decau amb la mida de la mostra a la velocitat òptima de $O(n^{-1/2})$, la dependència del paràmetre de confiança δ és exponencialment pitjor que a (2). Remetem a Catoni [8, proposició 6.2] per a una formulació precisa i per a un exemple senzill que (gairebé) satura la desigualtat de Txebixev.

Tot plegat condueix a una conclusió inevitable: si hom està buscant un estimador de la mitjana que sigui subgaussià per a qualsevol variable aleatòria que tingui una mitjana i una variància ben definides, llavors cal trobar alternatives a la mitjana mostral. Cal dir que, potser sorprenentment, existeixen estimadors de la mitjana que aconsegueixen un rendiment subgaussià per a totes les distribucions amb una variància finita. En les dues seccions següents es presenten i s'analitzen dos estimadors força diferents.

2.1 L'estimador mediana de les mitjanes

L'estimador mediana de les mitjanes que es presenta a continuació s'ha proposat de maneres diferents en diversos articles; vegeu Nemirovsky i Yudin [53], Hsu [28], Jerrum, Valiant i Vazirani [33] i Alon, Matias i Szegedy [1].

L'estimador mediana de les mitjanes divideix primer les dades en k grups de mida aproximadament igual, calcula la mitjana empírica de cada grup, i pren la mediana dels valors obtinguts.

Formalment, recordem que la mediana de k números reals $x_1, \dots, x_k \in \mathbb{R}$ es defineix com a $M(x_1, \dots, x_k) = x_i$, on x_i és tal que

$$|\{j \in [k] : x_j \le x_i\}| \ge \frac{k}{2} \quad \text{i} \quad |\{j \in [k] : x_j \ge x_i\}| \ge \frac{k}{2}.$$

(Si hi ha diversos índexs i que s'ajusten a la descripció anterior, prenem el més petit.)

Considerem ara $1 \le k \le n$ i dividim el conjunt $[n] = \{1, \dots, n\}$ en k blocs $B_1, \dots, B_k$, cadascun d'ells de mida $|B_i| \ge \lfloor n/k \rfloor \ge 2$.

Donat $X_1, \dots, X_n$, calculem la mitjana mostral de cada bloc

$$Z_j = \frac{1}{|B_j|} \sum_{i \in B_j} X_i$$

i definim l'estimador mediana de les mitjanes com a $\hat{\mu}_n = M(Z_1, \dots, Z_k)$.

Per copsar de manera intuïtiva per què funciona aquest estimador, observem que, per a cada bloc, la mitjana empírica és un estimador no esbiaixat de la mitjana, amb desviació estàndard $\sigma/\sqrt{n/k}$. Per tant, la mediana de la distribució de les mitjanes empíriques per blocs difereix com a màxim $\sigma/\sqrt{n/k}$ de l'esperança. Així, doncs, la mediana empírica és un estimador altament concentrat d'aquesta mediana.

A continuació establim una fita del rendiment de l'estimador. Per simplicitat, suposem que n és divisible per k, de manera que cada bloc té $m = n/k$ elements.

TEOREMA 2. *Siguin $X_1, \dots, X_n$ variables aleatòries independents i distribuïdes de manera idèntica amb mitjana μ i variància σ^2. Siguin m, k enters positius i suposem que $n = m\,k$. Aleshores, l'estimador mediana de mitjanes $\hat{\mu}_n$ amb k blocs satisfà*

$$\mathbb{P}\{|\hat{\mu}_n - \mu| > \sigma\sqrt{4/m}\} \le e^{-k/8}.$$

En particular, per a qualsevol $\delta \in (0,1)$, si $k = \lceil 8\log(1/\delta) \rceil$, aleshores, amb probabilitat almenys $1 - \delta$,

$$|\hat{\mu}_n - \mu| \le \sigma\sqrt{\frac{32\log(1/\delta)}{n}}.$$

PROVA. Per la desigualtat de Txebixev, per a cada $j = 1, \ldots, k$, amb probabilitat almenys $3/4$,

$$|Z_j - \mu| \leq \sigma\sqrt{\frac{4}{m}}.$$

Així, $|\hat{\mu}_n - \mu| > \sigma\sqrt{4/m}$ implica que almenys $k/2$ dels mitjanes Z_j són tals que $|Z_j - \mu| > \sigma\sqrt{4/m}$. Per tant, si $\mathrm{Bin}(k, 1/4)$ és una variable aleatòria binomial $(k, 1/4)$,

$$\begin{aligned}\mathbb{P}\{|\hat{\mu}_n - \mu| > \sigma\sqrt{4/m}\} &\leq \mathbb{P}\{\mathrm{Bin}(k, 1/4) \geq \frac{k}{2}\} = \\ &= \mathbb{P}\left\{\mathrm{Bin}(k, 1/4) - \mathbb{E}\,\mathrm{Bin}(k, 1/4) \geq \frac{k}{4}\right\} \leq \\ &\leq e^{-k/8},\end{aligned}$$

on en l'últim pas hem utilitzat una fita exponencial estàndard per a la cua d'una variable aleatòria binomial (per exemple, la desigualtat de Hoeffding [24]). □

El teorema 2 mostra que l'estimador mediana de les mitjanes té un rendiment subgaussià amb $L = 8$ per a totes les distribucions amb una variància finita. Tanmateix, és important assenyalar que l'estimador $\hat{\mu}_n$ depèn del nivell de confiança δ ja que el nombre de blocs k es tria en funció de δ. Es tracta d'una propietat poc desitjable, ja que per a diferents valors del paràmetre de confiança δ s'obté un estimador puntual diferent. Malgrat això, sense informació addicional sobre la distribució, és impossible construir un sol estimador que sigui subgaussià per a un rang no trivial de valors del paràmetre de confiança δ; vegeu Devroye, Lerasle, Lugosi i Oliveira [17] per a més detalls. No obstant, en el teorema 4, que enunciem més endavant, es presenta un resultat més positiu.

Els resultats de Bubeck, Cesa-Bianchi i Lugosi [7] i Devroye, Lerasle, Lugosi i Oliveira [17] mostren que l'estimador mediana de les mitjanes es pot utilitzar fins i tot si la distribució de X_i té una variància infinita però té un moment finit d'ordre $1 + \alpha$ per a alguna $\alpha \in (0, 1)$. En concret, tenim el resultat següent:

TEOREMA 3. *Sigui $\alpha \in (0, 1]$ i suposem que $X_1, \ldots, X_n$ són variables aleatòries independents i distribuïdes idènticament amb mitjana μ i moment central $(1+\alpha)$-èsim $M = \mathbb{E}[|X_i - \mu|^{1+\alpha}]$. Siguin m, k números enters positius i suposem que $n = m\,k$. Aleshores, l'estimador mediana de mitjanes amb $k = \lceil 8\log(2/\delta)\rceil$ blocs satisfà*

$$\mathbb{P}\left\{|\hat{\mu}_n - \mu| > 8\left(\frac{12M^{1/\alpha}\log(1/\delta)}{n}\right)^{\alpha/(1+\alpha)}\right\} \leq \delta.$$

A més, per a qualsevol estimador de la mitjana $\hat{\mu}_n$, existeix una distribució amb mitjana μ i moment central $(1 + \alpha)$-èsim M tal que

$$\mathbb{P}\left\{|\hat{\mu}_n - \mu| > \left(\frac{M^{1/\alpha}\log(2/\delta)}{n}\right)^{\alpha/(1+\alpha)}\right\} \geq \delta.$$

La prova de la primera part s'obté demostrant que, si $c(\alpha)$ és una constant adient que només depèn de α i

$$\eta \geq c(\alpha)(\mathbb{E}|X_i - \mu|^{1+\alpha})^{1/(1+\alpha)} \left(\frac{1}{m}\right)^{\alpha/(1+\alpha)},$$

aleshores

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \eta\right) \leq 0.2.$$

La prova de la segona afirmació va en la línia del teorema 1.

Acabem aquesta secció demostrant que, si la distribució de X té un moment finit d'ordre $2+\alpha$ per a algun $\alpha > 0$, aleshores l'estimador mediana de mitjanes té un rendiment subgaussià sota un rang més ampli d'eleccions del paràmetre k que compta el nombre de blocs. La fita següent es deu a Minsker i Strawn [52]. Per simplificar l'exposició, només considerem el cas $\alpha = 1$.

TEOREMA 4. *Siguin $X_1, \ldots, X_n$ variables aleatòries independents distribuïdes idènticament amb mitjana μ, variància σ^2, i tercer moment central $\rho = \exp|X - \mu|^3$. Siguin m, k números enters positius tals que $n = m\,k$. Suposem que*

$$\sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}} \leq 1/4. \tag{4}$$

Aleshores, l'estimador mediana de mitjanes $\hat{\mu}_n$ amb k blocs satisfà que, amb probabilitat almenys $1-\delta$,

$$|\hat{\mu}_n - \mu| \leq \frac{1}{c}\left(\sigma\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\rho k}{2\sigma^2 n}\right),$$

on $c = \phi(\Phi^{-1}(3/4))$ és una constant, mentre que ϕ i Φ denoten, respectivament, les funcions de densitat i de distribució normals estàndard.

Observem que el primer terme a la part dreta de la fita és de forma subgaussiana. El segon terme és menor que el primer sempre que el nombre k de blocs satisfaci

$$k \leq \frac{2\sigma^3}{\rho}\sqrt{n\log(2/\delta)}.$$

En particular, $k \leq \frac{2\sigma^3}{\rho}\sqrt{n}$ és suficient per tal d'obtenir un rendiment subgaussià. Això és bo, ja que amb aquesta opció l'estimador no depèn del valor del paràmetre de confiança δ i, simultàniament, l'estimador és subgaussià per a tot el rang de valors de δ permesos per la condició (4). A més, observem que es pot triar el nombre de blocs molt més gran que l'elecció suggerida pel teorema 2. En particular, k pot ser tan gran com un múltiple constant de $\sqrt{n}$. En aquest cas, l'estimador mediana de mitjanes és subgaussià simultàniament

per a tota $\delta \geq e^{-c_0\sqrt{n}}$ per a una constant c_0 adequada. El preu a pagar és la hipòtesi addicional de l'existència del tercer moment. Minsker i Strawn [52] també demostren que, quan $k = o(\sqrt{n})$, aleshores, sota les hipòtesis del teorema 4, $\sqrt{n}(\hat{\mu}_n - \mu)$ és asimptòticament normal amb mitjana zero i variància $\sigma^2\pi/2$.

PROVA. Observem que $\hat{\mu}_n \in [\mu - a, \mu + a]$ si $a > 0$ és tal que

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j-\mu\leq a} \geq \frac{1}{2} \quad \text{i} \quad \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j-\mu\geq -a} \geq \frac{1}{2}.$$

Demostrem que, amb probabilitat almenys $1-\delta$, podem prendre

$$a = \frac{1}{c}\left(\sigma\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\rho k}{2\sigma^2 n}\right).$$

A tal efecte, notem que

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j-\mu\leq a} = \frac{1}{k}\sum_{j=1}^{k} (\mathbb{1}_{Z_j-\mu\leq a} - \mathbb{P}\{Z_j-\mu\leq a\}) + \\ + \left(\mathbb{P}\{Z_1-\mu\leq a\} - \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\}\right) + \\ + \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\}\end{aligned}$$

(on G és una variable aleatòria normal estàndard).

Notem primer que, per la desigualtat de Hoeffding, amb probabilitat almenys $1-\delta/2$,

$$\frac{1}{k}\sum_{j=1}^{k} (\mathbb{1}_{Z_j-\mu\leq a} - \mathbb{P}\{Z_j-\mu\leq a\}) \geq -\sqrt{\frac{\log(2/\delta)}{2k}}.$$

Per al segon terme del costat dret, podem utilitzar el teorema de Berry-Esseen (vegeu Shevtsova [56]), que implica que

$$\mathbb{P}\{Z_1-\mu\leq a\} - \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\} \geq -\frac{\rho}{2\sigma^3\sqrt{m}}.$$

Per tant, tenim que, amb probabilitat almenys $1-\delta/2$,

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j-\mu\leq a} \geq \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\} - \sqrt{\frac{\log(2/\delta)}{2k}} - \frac{\rho}{2\sigma^3\sqrt{m}}.$$

Així, $(1/k)\sum_{j=1}^{k} \mathbb{1}_{Z_j-\mu\leq a} \geq \frac{1}{2}$ amb probabilitat almenys $1-\delta/2$, sempre que a sigui tal que

$$\mathbb{P}\left\{G \leq a\frac{\sqrt{m}}{\sigma}\right\} \geq \frac{1}{2} + \sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}}.$$

Si $\sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}} \le 1/4$, aleshores n'hi ha prou de considerar valors de a tals que $a\sqrt{m}/\sigma \le \Phi^{-1}(3/4)$. Llavors,

$$\mathbb{P}\left\{G \le a\frac{\sqrt{m}}{\sigma}\right\} \ge \frac{1}{2} + c\frac{a\sqrt{m}}{\sigma}$$

amb $c = \phi(\Phi^{-1}(3/4))$. Per tant, podem prendre

$$a = \frac{\sigma}{c\sqrt{m}}\left(\sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}}\right) = \frac{1}{c}\left(\sigma\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\rho k}{2\sigma^2 n}\right).$$

El mateix argument mostra que, amb probabilitat almenys $1 - \delta/2$,

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j - \mu \ge -a} \ge \frac{1}{2}$$

per l'elecció de a que hem fet més amunt. □

2.2 L'estimador de Catoni

A continuació, presentem un enfocament completament diferent per construir un estimador de mitjanes, introduït i analitzat per Catoni [8]. Per introduir la idea de Catoni, observem primer que la mitjana empírica $\overline{\mu}_n$ és precisament la solució $y \in \mathbb{R}$ de l'equació

$$\sum_{i=1}^{n}(X_i - y) = 0.$$

Catoni va proposar substituir el costat esquerre de l'equació anterior per una altra funció estrictament decreixent de y de la forma

$$R_{n,\alpha}(y) = \sum_{i=1}^{n}\psi(\alpha(X_i - y)),$$

on $\psi\colon \mathbb{R} \to \mathbb{R}$ és una funció creixent antisimètrica i $\alpha \in \mathbb{R}$ és un paràmetre. La idea és que, si $\psi(x)$ augmenta molt més lentament que x, llavors es redueix l'efecte de les dades atípiques presents a causa de les cues pesades. Catoni ofereix tot un ventall de *funcions d'influència* ψ. Per facilitar l'exposició, triem una funció concreta, com ara

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2) & \text{if } x \ge 0, \\ -\log(1 - x + x^2/2) & \text{if } x < 0. \end{cases}$$

Definim l'estimador de la mitjana de Catoni $\widehat{\mu}_{\alpha,n}$ com el valor únic y tal que $R_{n,\alpha}(y) = 0$ amb aquesta elecció de ψ. Com que $\psi(x) \leq \log(1 + x + x^2/2)$ per a tot $x \in \mathbb{R}$, tenim que, per a tot $y \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}[e^{R_{n,\alpha}(y)}] &\leq \left(\mathbb{E}\left[1 + \alpha(X - y) + \frac{\alpha^2(X-y)^2}{2}\right]\right)^n = \\
&= \left(1 + \alpha(\mu - y) + \frac{\alpha^2(\sigma^2 + (\mu - y)^2)}{2}\right)^n \leq \\
&\leq \exp\left(n\alpha(\mu - y) + \frac{n\alpha^2(\sigma^2 + (\mu - y)^2)}{2}\right),
\end{aligned}$$

sempre que els X_i tinguin una variància finita σ^2. Així, per la desigualtat de Markov, tenim que, per a qualsevol $y \in \mathbb{R}$ fix i $\delta \in (0,1)$,

$$\mathbb{P}\left\{R_{n,\alpha}(y) \geq n\alpha(\mu - y) + \frac{n\alpha^2(\sigma^2 + (\mu - y)^2)}{2} + \log(1/\delta)\right\} \leq \delta.$$

Suposem que el paràmetre α és tal que $\alpha^2\sigma^2 + 2\log(1/\delta)/n \leq 1$. Aleshores el polinomi quadràtic en y

$$n\alpha(\mu - y) + \frac{n\alpha^2(\sigma^2 + (\mu - y)^2)}{2} + \log(1/\delta)$$

té almenys una arrel. En particular, prenent l'arrel menor

$$y_+ = \mu + \frac{\frac{\alpha\sigma^2}{2} + \frac{\log(1/\delta)}{n\alpha}}{\frac{1}{2} + \frac{1}{2}\sqrt{1 - \alpha^2\sigma^2 - \frac{2\log(1/\delta)}{n}}},$$

tenim que $R_{n,\alpha}(y_+) < 0$ amb probabilitat almenys $1 - \delta$. Com que $R_{n,\alpha}(y)$ és estrictament decreixent, això implica que $\widehat{\mu}_{\alpha,n} < y_+$ amb probabilitat almenys $1 - \delta$. Un argument simètric serveix per a demostrar que $\widehat{\mu}_{\alpha,n} > y_-$ amb probabilitat almenys $1 - \delta$, on

$$y_- = \mu - \frac{\frac{\alpha\sigma^2}{2} + \frac{\log(1/\delta)}{n\alpha}}{\frac{1}{2} + \frac{1}{2}\sqrt{1 - \alpha^2\sigma^2 - \frac{2\log(1/\delta)}{n}}}.$$

Ara, mitjançant fites senzilles i escollint el paràmetre α per tal d'optimitzar les fites, obtenim la següent estimació del rendiment.

TEOREMA 5. *Siguin $X_1, \ldots, X_n$ variables aleatòries independents i distribuïdes de manera idèntica amb mitjana μ i variància σ^2. Sigui $\delta \in (0,1)$ tal que $n > 2\log(1/\delta)$. L'estimador de la mitjana de Catoni $\hat{\mu}_{n,\alpha}$ amb paràmetre*

$$\alpha = \sqrt{\frac{2\log(1/\delta)}{n\sigma^2\left(1 + \frac{2\log(1/\delta)}{n - 2\log(1/\delta)}\right)}}$$

satisfà, amb probabilitat almenys $1 - 2\delta$,

$$|\hat{\mu}_{n,\alpha} - \mu| < \sqrt{\frac{2\sigma^2\log(1/\delta)}{n - 2\log(1/\delta)}}. \tag{5}$$

El teorema destaca que, amb un paràmetre α triat de manera escaient, l'estimador de Catoni té un rendiment subgaussià. Cal remarcar que la constant $\sqrt{2}$ és la millor possible. Un desavantatge de l'estimador de Catoni respecte de la mediana de les mitjanes és que —almenys en la forma donada al teorema— depèn de la variància σ^2. En general, no és realista suposar el coneixement de σ^2. Si hom substitueix σ^2 a la fórmula de α per una fita superior v, llavors la fita (5) encara es manté si substituïm σ^2 per v. En cas que no existeixi cap fita superior bona per a σ^2, Catoni ([8]) mostra com seleccionar α a partir de les dades que tenen un rendiment gairebé òptim. Huber ([30]) combina l'estimador mediana de les mitjanes amb l'estimador de Catoni en un procediment de dos passos i obté un estimador amb la constant del terme dominant òptima en la fita subgaussiana quan $|\sigma/\mu|$ està fitada per una constant coneguda.

Un altre problema —compartit amb l'estimador mediana de les mitjanes— és que l'estimador de Catoni també depèn del nivell de confiança requerit δ. Aquesta dependència, tal com s'ha esmentat anteriorment, és necessària. Una solució ràpida és utilitzar l'estimador amb un paràmetre independent de δ, tot i que l'estimació resultant, naturalment, no podrà ser subgaussiana. Una opció raonable és $\alpha = \sqrt{2/(n\sigma^2)}$. En aquest cas, és fàcil veure que, sempre que $n > 2(1 + \log(1/\delta))$, l'estimador de Catoni satisfà, amb probabilitat almenys $1 - 2\delta$,

$$|\hat{\mu}_{n,\alpha} - \mu| < \sqrt{\frac{\sigma^2}{2n}} \cdot \frac{1 + \log(1/\delta)}{1 - \frac{1+\log(1/\delta)}{n}}.$$

A causa d'un factor addicional $\sqrt{\log(1/\delta)}$, aquesta fita no és subgaussiana, però les probabilitats «subexponencials» de les cues són encara no trivials i útils.

2.3 Mitjana retallada

L'intent més natural de millorar el rendiment de la mitjana empírica és potser el d'eliminar possibles valors atípics mitjançant un truncament de X. De fet, l'estimador anomenat *mitjana retallada* (o *mitjana truncada*) es defineix eliminant una fracció de la mostra, que consisteix en els ϵn punts més grans i més petits per a algun paràmetre $\epsilon \in (0,1)$, i després fer la mitjana sobre la resta.

Aquesta idea és una de les eines més clàssiques en estadística robusta i remetem a Tukey i McLaughlin [60], Huber i Ronchetti [32], Bickel [3] i Stigler [58] per consultar els primers treballs sobre les propietats teòriques de l'estimador mitjana retallada. No obstant això, fa ben poc que es va introduir la propietat subgaussiana no asimptòtica de la mitjana retallada. En efecte, Oliveira i Orenstein ([54]) van demostrar que, si s'escull ϵ proporcionalment a $\log(1/\delta)/n$, aleshores l'estimador mitjana retallada té un rendiment subgaussià per a totes les distribucions amb una variància finita.

Per mostrar de la manera més senzilla com funciona aquesta idea, a continuació analitzem una variant senzilla de l'estimador mitjana retallada.

L'estimador divideix les dades en dues parts iguals. Una meitat s'utilitza per a determinar el nivell de truncament correcte. Amb els punts de l'altra meitat se'n calcula la mitjana, sense tenir en compte les dades que queden fora de la regió de truncament. Per comoditat de notació, suposem que les dades consisteixen en $2n$ còpies independents de la variable aleatòria X, denotades per $X_1, \dots, X_n, Y_1, \dots, Y_n$.

Per a $\alpha \le \beta$, definim la funció de truncament

$$\phi_{\alpha,\beta}(x) = \begin{cases} \beta & \text{si } x > \beta, \\ x & \text{si } x \in [\alpha, \beta], \\ \alpha & \text{si } x < \alpha, \end{cases}$$

i per a $x_1, \dots, x_m \in \mathbb{R}$ sigui $x_1^* \le x_2^* \le \cdots \le x_m^*$ la seva reordenació no decreixent.

Un cop establerta aquesta notació, la definició de l'estimador és la següent:

(1) Donat el nivell de confiança $\delta \ge 8e^{-3n/16}$, definim

$$\varepsilon = \frac{16\log(8/\delta)}{3n}.$$

(2) Siguin $\alpha = Y^*_{\varepsilon n}$ i $\beta = Y^*_{(1-\varepsilon)n}$ (assumint, per simplificar, que εn és un enter) i definim

$$\widehat{\mu}_{2n} = \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha,\beta}(X_i).$$

TEOREMA 6. *Siguin $X_1, \dots, X_n, Y_1, \dots, Y_n$ variables aleatòries independents i distribuïdes de manera idèntica amb mitjana μ i variància σ^2. Sigui $\delta \in (0,1)$ tal que $n > (16/3)\log(8/\delta)$. Aleshores, amb probabilitat almenys $1-\delta$,*

$$|\widehat{\mu}_{2n} - \mu| \le 9\sigma\sqrt{\frac{\log(8/\delta)}{n}}.$$

PROVA. Comencem demostrant que el nivell de truncament és proper als quantils adequats de la distribució. Amb aquesta finalitat, per a $p \in (0,1)$, introduïm els quantils

$$Q_p = \sup\{M \in \mathbb{R} : \mathbb{P}\{X \ge M\} \ge 1-p\}.$$

Per facilitar l'exposició, suposem que X té una distribució no atòmica. (Aquesta suposició no és necessària, però simplifica la notació.) En aquest cas, $\mathbb{P}\{X > Q_p\} = \mathbb{P}\{X \geq Q_p\} = 1 - p$.

Mitjançant una aplicació senzilla d'una desigualtat de cua binomial (en particular, la desigualtat de Bernstein), amb probabilitat almenys $1 - 2\exp(-(3/16)\varepsilon n)$, tenim tant que

$$|\{i \in [n] : Y_i \geq Q_{1-2\varepsilon}\}| \geq \varepsilon n$$

com que

$$|\{i \in [n] : Y_i \leq Q_{1-\varepsilon/2}\}| \geq (1-\varepsilon)n.$$

Això implica que, amb probabilitat almenys $1 - 2\exp(-(3/16)\varepsilon n)$,

$$Q_{1-2\varepsilon} \leq Y^*_{(1-\varepsilon)n} \leq Q_{1-\varepsilon/2}. \tag{6}$$

Seguint el mateix argument, amb probabilitat almenys $1-2\exp(-(3/16)\varepsilon n)$,

$$Q_{\varepsilon/2} \leq Y^*_{\varepsilon n} \leq Q_{2\varepsilon}. \tag{7}$$

A partir d'aquí, simplement necessitem provar que $|\mathbb{E}\phi_{\alpha,\beta}(X) - \mu|$ és petit i que $(1/n)\sum_{i=1}^n \phi_{\alpha,\beta}(X_i)$ es concentra al voltant de la seva mitjana.

Per al primer pas, considerem l'esdeveniment E consistent en el fet que tant (6) com (7) són certes. Aquest esdeveniment té una probabilitat d'almenys $1 - 4\exp(-(3/16)\varepsilon n) = 1 - \delta/2$. En l'esdeveniment E,

$$\begin{aligned}
&|\mathbb{E}[\phi_{\alpha,\beta}(X)|Y_1,\dots,Y_n] - \mu| \leq \\
&\quad \leq |\mathbb{E}[(X-\alpha)\mathbb{1}_{X\leq\alpha}|Y_1,\dots,Y_n]| + |\mathbb{E}[(X-\beta)\mathbb{1}_{X\geq\beta}|Y_1,\dots,Y_n]| \leq \\
&\quad \leq |\mathbb{E}(X-Q_{2\varepsilon})\mathbb{1}_{X\leq Q_{2\varepsilon}}| + |\mathbb{E}(X-Q_{1-2\varepsilon})\mathbb{1}_{X\geq Q_{1-2\varepsilon}}|.
\end{aligned}$$

Per a relacionar aquests dos termes, cal tenir en compte que, per la desigualtat de Txebixev,

$$2\varepsilon = \mathbb{P}\{X \geq Q_{1-2\varepsilon}\} \leq \frac{\sigma_X^2}{(Q_{1-2\varepsilon}-\mu)^2},$$

i, en particular,

$$Q_{1-2\varepsilon} \leq \mu + \frac{\sigma}{\sqrt{2\varepsilon}}.$$

Per tant, per la desigualtat de Cauchy-Schwarz,

$$\begin{aligned}
|\mathbb{E}(X-Q_{1-2\varepsilon})\mathbb{1}_{X\geq Q_{1-2\varepsilon}}| &= |\mathbb{E}(X-\mu) - (Q_{1-2\varepsilon}-\mu))\mathbb{1}_{X\geq Q_{1-2\varepsilon}}| \leq \\
&\leq \mathbb{E}|(X-\mu)|\mathbb{1}_{X\geq Q_{1-2\varepsilon}} + (Q_{1-2\varepsilon}-\mu)\mathbb{P}\{X \geq Q_{1-2\varepsilon}\} \leq \\
&\leq \sigma\sqrt{\mathbb{P}\{X \geq Q_{1-2\varepsilon}\}} + 2\varepsilon(Q_{1-2\varepsilon}-\mu) \leq \\
&\leq \sigma\sqrt{8\varepsilon}.
\end{aligned}$$

Un argument simètric prova que $|\mathbb{E}(X - Q_{2\varepsilon})\mathbb{1}_{X \le Q_{2\varepsilon}}| \le \sigma\sqrt{8\varepsilon}$, i per tant, en l'esdeveniment E, tenim que

$$|\mathbb{E}[\phi_{\alpha,\beta}(X)|Y_1,\dots,Y_n] - \mu| \le \sigma\sqrt{32\varepsilon} \le 6\sigma\sqrt{\frac{\log(8/\delta)}{n}}$$

per a la nostra elecció de ϵ. A continuació, considerem

$$Z = \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha,\beta}(X_i) - \mathbb{E}[\phi_{\alpha,\beta}(X)|Y_1,\dots,Y_n]$$

i observem que

$$Z = \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha-\mu,\beta-\mu}(X_i - \mu) - \mathbb{E}[\phi_{\alpha-\mu,\beta-\mu}(X - \mu)|Y_1,\dots,Y_n].$$

Per tant, en l'esdeveniment E (que només depèn de $Y_1,\dots,Y_n$), Z és una mitjana de variables aleatòries centrades que està fitada punt a punt per $M = \max\{|Q_{\varepsilon/2} - \mu|, |Q_{1-\varepsilon/2} - \mu|\} \le \sigma\sqrt{2/\varepsilon}$ i la variància de la qual és com a màxim σ^2. Per tant, per la desigualtat de Bernstein, amb probabilitat almenys $1 - \delta/2$,

$$Z \le \sigma\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{\log(2/\delta)\sigma\sqrt{2/\varepsilon}}{n} \le 3\sigma\sqrt{\frac{\log(2/\delta)}{n}}.$$

Ajuntant totes les peces, obtenim la fita anunciada. □

A més de la seva senzillesa conceptual, un avantatge important de la mitjana retallada en comparació amb altres estimadors amb rendiment subgaussià és que és robusta a la contaminació antagonista de les dades. Aquesta propietat es va formalitzar i demostrar a [47], on també se'n presenta i analitza una extensió multivariant.

Tanquem aquesta secció observant que, en un treball recent, Lee i Valiant ([41]) construeixen un estimador subgaussià amb la constant (gairebé) òptima $L = \sqrt{2} + o(1)$. El seu estimador es basa en una combinació intel·ligent de mediana de les mitjanes, mitjana retallada i estimador de Catoni. Per una altra banda, Minsker i Ndaoud ([51]) van proposar un enfocament diferent. Igual que la mediana de mitjanes, el seu estimador de la mitjana també comença calculant mitjanes empíriques en blocs disjunts de les dades. A continuació, tornen a ponderar les mitjanes del bloc en funció de la seva desviació estàndard empírica. Utilitzant propietats no trivials de sumes autonormalitzades, obtenen un estimador que no només és subgaussià sinó que també és asimptòticament eficient, en el sentit que l'estimador és asimptòticament normal amb una variància asimptòtica tan petita com sigui possible en el sentit minimax.

3 Estimació de la mitjana d'un vector aleatori

A continuació, discutim les extensions del problema d'estimació de la mitjana al cas multivariant. Per tal d'establir el problema, considerem un vector aleatori X que pren valors a $\mathbb{R}^d$. Suposem que el vector de mitjanes $\mu = \mathbb{E}X$ i la matriu de covariància $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$ existeixen. Donades n mostres independents i distribuïdes de manera idèntica $X_1, \dots, X_n$ extretes de la distribució de X, volem estimar el vector de mitjanes.

Igual que en el cas univariant, una elecció natural és la mitjana mostral $\overline{\mu}_n = (1/n)\sum_{i=1}^n X_i$, que té un comportament gairebé òptim sempre que la distribució tingui una cua prou lleugera. Tanmateix, com passa en el cas univariant, sempre que ens preocupin les cues pesades, cal evitar la mitjana mostral ja que pot tenir un rendiment subòptim.

3.1 Rendiment subgaussià

Hem vist diversos exemples d'estimadors de mitjanes amb un rendiment subgaussià. Per tal d'establir correctament el nostre objectiu per al cas d-dimensional, primer hem d'entendre què vol dir «rendiment subgaussià». Com en el cas univariant, hom voldria construir estimadors que fossin «a prop» de la veritable mitjana μ amb «alta probabilitat». A partir d'ara, $\|\cdot\|$ denotarà la norma euclidiana habitual a $\mathbb{R}^d$.

Si X té una distribució normal multivariant amb vector de mitjanes μ i matriu de covariància Σ, llavors la mitjana mostral $\overline{\mu}_n$ també és normal multivariant amb mitjana μ i matriu de covariància $(1/n)\Sigma$. Així, per a tot $t > 0$,

$$\mathbb{P}\{\|\overline{\mu}_n - \mu\| \geq \mathbb{E}\|\overline{\mu}_n - \mu\| + t\} = \mathbb{P}\{\|\overline{X}\| - \mathbb{E}\|\overline{X}\| \geq t\sqrt{n}\},$$

on $\overline{X}$ és un vector gaussià en $\mathbb{R}^d$ amb mitjana zero i matriu de covariància Σ. Una propietat clau dels vectors gaussians és que $\overline{X}$ té la mateixa distribució que $\Sigma^{1/2}Y$, on Y és un vector normal estàndard (és a dir, amb una matriu de covariància igual a la identitat i amb mitjana zero) i $\Sigma^{1/2}$ és l'arrel quadrada semidefinida positiva de Σ. Observem també que per a tots els $y, y' \in \mathbb{R}^d$,

$$\Big| \|\Sigma^{1/2}y\| - \|\Sigma^{1/2}y'\| \Big| \leq \|\Sigma^{1/2}(y-y')\| \leq \|\Sigma^{1/2}\|_{2\to 2} \cdot \|y - y'\|,$$

on $\|\Sigma^{1/2}\|_{2\to 2}$ és la norma espectral de $\Sigma^{1/2}$. Així, $\Sigma^{1/2}y$ és una funció de Lipschitz de $y \in \mathbb{R}^d$ amb una constant de Lipschitz $\|\Sigma^{1/2}\|_{2\to 2} = \sqrt{\lambda_{\max}}$, on $\lambda_{\max} = \lambda_{\max}(\Sigma)$ denota el valor propi més gran de la matriu de covariància Σ. Ara, de la desigualtat de concentració gaussiana de Tsirelson, Ibragimov i Sudakov [14] (vegeu també Ledoux [39] i Boucheron, Lugosi i Massart [6] per a més informació), es dedueix que

$$\mathbb{P}\{\|\overline{X}\| - \mathbb{E}\|\overline{X}\| \geq t\sqrt{n}\} \leq e^{-nt^2/(2\lambda_{\max})}.$$

Observant que

$$\mathbb{E}\|\overline{X}\| \leq \sqrt{\mathbb{E}\|\overline{X}\|^2} = \sqrt{\mathrm{Tr}(\Sigma)},$$

on $\mathrm{Tr}(\Sigma)$ és la traça de la matriu de covariància Σ, tenim que, per a $\delta \in (0,1)$, amb probabilitat almenys $1-\delta$,

$$\|\overline{\mu}_n - \mu\| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}}. \tag{8}$$

Així, en el cas multivariant, direm que un estimador de la mitjana és *subgaussià* si, amb probabilitat almenys $1-\delta$, satisfà una desigualtat de la forma (8) (amb factors constants possiblement diferents). Observem que, per a qualsevol distribució amb mitjana μ i matriu de covariància Σ, l'error quadràtic mitjà de la mitjana empírica és igual a

$$\mathbb{E}\|\overline{\mu}_n - \mu\|^2 = \frac{\mathrm{Tr}(\Sigma)}{n}.$$

En particular, $\mathbb{E}\|\overline{\mu}_n - \mu\| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$. Una característica important de la propietat subgaussiana (8) és que les fluctuacions aleatòries estan controlades per la norma espectral $\lambda_{\max}$ de la matriu de covariància, que possiblement és molt més petita que $\mathrm{Tr}(\Sigma)$, la suma de tots els valors propis de Σ.

3.2 Mediana de mitjanes multivariant

Per a distribucions no gaussianes i possiblement de cua pesada, no es pot esperar un comportament subgaussià de la mitjana mostral semblant a (8).

Com a alternativa, es pot intentar estendre l'estimador mediana de mitjanes al cas multivariant. Una idea òbvia és dividir les dades en blocs disjunts, calcular la mitjana empírica dins de cada bloc i calcular una mediana multivariant de les mitjanes empíriques obtingudes. Tanmateix, no hi ha una noció estàndard de mediana per a dades multivariants, i no està del tot clar quina definició de mediana multivariant funciona millor per als estimadors medianes de mitjanes. Entre les nombroses possibilitats, esmentem la *mediana per coordenades*, la *mediana geomètrica (o espacial)*, la *mediana de Tukey (o de semiespai)*, la *mediana d'Oja* i la *mediana de Liu*; vegeu la revisió de Small [57], que conté, a més, referències rellevants.

Independentment de quina noció de mediana multivariant decidim adoptar, comencem dividint $[n] = \{1,\dots,n\}$ en k blocs $B_1,\dots,B_k$, cadascun dels quals de mida $|B_i| \geq \lfloor n/k \rfloor \geq 2$. Aquí, k és un paràmetre de l'estimador que es triarà més endavant. Per simplificar, suposem que $k\,m = n$ per a algun enter positiu m. Com abans, calculem la mitjana mostral dels vectors aleatoris dins de cada bloc: per a $j = 1,\dots,k$, considerem

$$Z_j = \frac{1}{m}\sum_{i\in B_j} X_i.$$

El primer intent més natural és, possiblement, definir $\hat{\mu}_n$ com el vector de les medianes coordenada a coordenada de les Z_j (és a dir, la ℓ-èsima component del vector $\hat{\mu}_n$ és la mediana de les ℓ-èsimes components de $Z_1, \dots, Z_k$, per a $\ell \in [d]$). Aleshores, el teorema 2 i la fita de la unió (desigualtat de Boole) impliquen que, per a qualsevol $\delta \in (0,1)$, prenent $k = \lceil 8 \log(1/\delta) \rceil$, amb probabilitat almenys $1 - \delta$,

$$\|\hat{\mu}_n - \mu\| \le \sqrt{\frac{32 \operatorname{Tr}(\Sigma) \log(d/\delta)}{n}},$$

on hem utilitzat el fet que $\operatorname{Tr}(\Sigma) = \mathbb{E}\|X - \mathbb{E}X\|^2$ és la suma de les variàncies de les d components de X. Clarament, aquesta fita queda lluny de la desigualtat subgaussiana (8) per diverses raons. En primer lloc, no és independent de la dimensió ja que d hi apareix explícitament. Segurament, però, és més important encara que $\log(d/\delta)$ ve multiplicat per $\operatorname{Tr}(\Sigma)$ en lloc de $\lambda_{\max}(\Sigma)$, i això pot ser una diferència rellevant en problemes d'alta dimensió, especialment quan hom està interessat en probabilitats d'error petites. Un exemple il·lustratiu és quan tots els valors propis de Σ són idèntics i iguals a $\lambda_{\max}$. Si la dimensió d és gran, l'expressió (8) és de l'ordre de $\sqrt{(\lambda_{\max}/n)(d + \log(1/\delta))}$ mentre que la fita anterior només dona l'ordre $\sqrt{(\lambda_{\max}/n)(d \log(d/\delta))}$.

De tota manera, això es pot millorar amb força facilitat amb una noció diferent (no estàndard) de mediana a la definició de l'estimació: triem $\hat{\mu}_n$ com el punt de $\mathbb{R}^d$ que té la propietat que la bola euclidiana centrada en $\hat{\mu}_n$ que conté més de $k/2$ dels punts Z_j té un radi mínim. Com que $\mathbb{E}\|Z_j - \mu\|^2 = \operatorname{Tr}(\Sigma)/m$, per la desigualtat de Txebixev, $\|Z_j - \mu\| \le r := 2\sqrt{\operatorname{Tr}(\Sigma)/m}$ amb una probabilitat d'almenys $3/4$. Així, escollint $k = \lceil 8 \log(1/\delta) \rceil$, tenim que, amb probabilitat almenys $1 - \delta$, més de la meitat dels punts Z_j compleixen

$$\|Z_j - \mu\| \le r.$$

Denotem ara aquest esdeveniment amb E (així, $\mathbb{P}\{E\} \ge 1 - \delta$.) En l'esdeveniment E, aquest radi és com a màxim r. Per tant, almenys un dels Z_j es troba a la distància r tant de μ com de $\hat{\mu}_n$ i, per la desigualtat triangular, $\|\hat{\mu}_n - \mu\| \le 2r$. Hem obtingut, doncs, la proposició següent.

Proposició 7. *Siguin $X_1, \dots, X_n$ vectors aleatoris i.i.d. en $\mathbb{R}^d$ amb mitjana μ i matriu de covariància Σ. Sigui $\delta \in (0,1)$ i sigui $\hat{\mu}_n$ l'estimador definit anteriorment amb $k = \lceil 8 \log(1/\delta) \rceil$. Aleshores, amb probabilitat almenys $1 - \delta$,*

$$\|\hat{\mu}_n - \mu\| \le 4\sqrt{\frac{\operatorname{Tr}(\Sigma)(8 \log(1/\delta) + 1)}{n}}.$$

La fita de la proposició 7 és remarcable ja que no depèn de la dimensió i no es fa cap hipòtesi que no sigui l'existència de la matriu de covariància. Tanmateix, encara no assoleix una fita de rendiment subgaussià que s'assembli a (8). A més, la noció de mediana utilitzada aquí (és a dir, el centre de la bola més petita que conté almenys la meitat dels punts) és problemàtica des del

punt de vista computacional, ja que calcular aquesta mediana és un problema no trivial.

Una versió calculable de manera eficient d'una mediana multivariant és l'anomenada *mediana geomètrica*, definida com a

$$\widehat{\mu}_n = \operatorname*{argmin}_{m \in \mathbb{R}^d} \sum_{j=1}^{k} \|Z_i - m\|.$$

Aquest estimador va ser proposat per Minsker [50] i independentment per Hsu i Sabato [29] (vegeu també Lerasle i Oliveira [43]). En particular, Minsker ([50]) va demostrar que aquesta versió multivariant de l'estimador mediana de les mitjanes assoleix un rendiment semblant a la proposició 7. A més, calcular la mediana geomètrica i, per tant, l'estimador mediana de les mitjanes multivariant, implica resoldre un problema d'optimització convex. Així, la mediana geomètrica es pot aproximar de manera eficient; vegeu Cohen, Lee, Miller, Pachocki i Sidford [15] per al resultat més recent i per a la rica història del problema. Remetem el lector a Aloupis [2] per a una revisió dels aspectes computacionals d'altres nocions diverses de medianes multivariants.

Per a un estimador de la mitjana molt diferent, basat en la idea de la mediana de les mitjanes amb garanties «gairebé» subgaussianes però amb un cost computacional important, vegeu Joly, Lugosi i Oliveira [34].

Per tal d'aconseguir un rendiment realment subgaussià, hem de definir un nou estimador. A continuació, en definim un que aconsegueix el rendiment desitjat: aquest estimador es va introduir a [46], basat en la idea de torneigs de mediana de mitjanes. També esmentem un altre estimador, proposat a [47], que es defineix a partir d'una extensió multivariant de l'estimador mitjana retallada. Tanmateix, abans de presentar aquestes estimacions, recordem un estimador molt diferent introduït per Catoni i Giulini [10].

3.3 Llindar de la norma: l'estimador de Catoni-Giulini

En aquesta secció discutim breument un estimador notablement senzill, suggerit i analitzat per Catoni i Giulini [10]. L'estimador de Catoni-Giulini és

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \min\left(1, \frac{1}{\alpha \|X_i\|}\right), \tag{9}$$

on $\alpha > 0$ és un paràmetre (petit). Fixem-nos que $\widehat{\mu}_n$ és simplement una mitjana empírica dels X_i, on els punts que tenen una norma gran es fan tendir a zero. Aquesta estimació és trivial de calcular, a diferència dels estimadors més complexos que discutirem a la subsecció 3.4. D'altra banda, val a dir que fer tendir a zero és una mica arbitrari i antinatural. De fet, l'estimador no és invariant per translacions de les dades en el sentit que $\widehat{\mu}_n(X_1 + a, \ldots, X_n + a)$ no és necessàriament igual a $\widehat{\mu}_n(X_1, \ldots, X_n) + a$ quan $a \neq 0$.

Catoni i Giulini demostren que, si es tria el paràmetre com a

$$\alpha = \sqrt{\frac{c \log(1/\delta)}{v n}},$$

on $v \geq \lambda_{\max}$ i $c > 0$ és una constant numèrica, aleshores l'estimador (9) compleix, amb probabilitat almenys $1 - \delta$,

$$\|\hat{\mu}_n - \mu\| \leq C \sqrt{\frac{(\mathrm{Tr}(\Sigma) + v + \|\mu\|^2) \log(1/\delta)}{n}},$$

on C és una constant que només depèn de c. Aquesta fita és similar però més feble que la de la proposició 7, principalment a causa de dos fets. En primer lloc, l'estimador requereix un coneixement previ (una bona fita superior) de $\lambda_{\max}$, mentre que l'estimador mediana geomètrica de les mitjanes no assumeix aquesta informació prèvia. En segon lloc, $\|\mu\|^2$ apareix a la fita superior que, *a priori*, pot ser arbitràriament gran en comparació amb $\mathrm{Tr}(\Sigma)$. La presència d'aquest terme es deu a la manca d'invariància per translacions de l'estimador. Aquest segon problema es pot solucionar mitjançant la definició d'un estimador de dues etapes: primer es pot utilitzar un estimador invariant per translacions com la mediana geomètrica de mitjanes definit a la secció anterior per tal d'obtenir una estimació aproximada de la mitjana. Aleshores, utilitzant un nou conjunt de dades independents, es poden centrar les dades a la mitjana estimada i després utilitzar l'estimador de Catoni-Giulini per a les dades centrades. Aquest nou estimador és invariant per translacions, i el terme $\|\mu\|^2$ es pot substituir per l'error quadràtic de l'estimador del primer pas, és a dir, per $\mathrm{Tr}(\Sigma) \log(1/\delta)/n$. Però, fins i tot amb aquesta modificació, la fita no és subgaussiana en el sentit de (8).

Tanmateix, és remarcable que el rendiment de l'estimador de Catoni-Giulini s'acosti a ser subgaussià en el sentit desitjat amb només una petita hipòtesi addicional. En particular, si $\mathbb{E}\|X\|^\beta < \infty$ per a alguns $\beta > 2$, llavors, amb la mateixa elecció de α que l'anterior, tenim que

$$\|\hat{\mu}_n - \mu\| \leq C \left(\sqrt{\frac{v \log(1/\delta)}{n}} + \sqrt{\frac{(\mathrm{Tr}(\Sigma) + v)}{n}} + \frac{\kappa_\beta}{n^{(\beta-1)/2}} \right),$$

on κ_β és una constant que depèn de β i del β-èsim moment de $\|X\|$. Així, si el paràmetre anterior v és una bona estimació de $\lambda_{\max}$ en el sentit que està fitat per un múltiple constant, aleshores els dos primers termes de la fita són de la forma subgaussiana desitjada. El tercer terme és d'ordre més petit, encara que, novament, pot ser arbitràriament gran si la mitjana està lluny de l'origen, cosa que es pot solucionar fent l'estimador més complex. Remetem a Catoni i Giulini [9] per a altres estimacions amb un esperit similar i per a més discussió sobre la qüestió. Les tècniques de demostració de [9, 10] es basen en les anomenades desigualtats *PAC-bayesianes*, els detalls de les quals sobrepassen l'abast d'aquesta revisió.

3.4 Torneigs de mediana de mitjanes

En aquesta secció introduïm un estimador de la mitjana, proposat per Lugosi i Mendelson [46], amb un rendiment subgaussià per a totes les distribucions la matriu de covariància de les quals existeixi. L'estimador que es presenta a continuació és l'exemple primari i més senzill del que anomenem *torneigs de mediana de mitjanes*.

Recordem que partim d'una mostra $X_1, \ldots, X_n$ de vectors aleatoris i.i.d. a $\mathbb{R}^d$. Com en el cas de l'estimador mediana de les mitjanes, comencem dividint el conjunt $\{1, \ldots, n\}$ en k blocs $B_1, \ldots, B_k$, cadascun dels quals de mida $|B_j| \geq m = \lfloor n/k \rfloor$, on k és un paràmetre de l'estimador el valor del qual depèn del nivell de confiança desitjat, tal com s'especifica a continuació. Per simplificar la presentació, suposem que n és divisible per k i, per tant, $|B_j| = m$ per a tot $j = 1, \ldots, k$.

Definim la mitjana mostral dins de cada bloc com a

$$Z_j = \frac{1}{m} \sum_{i \in B_j} X_i.$$

Per a cada $a \in \mathbb{R}^d$, sigui

$$T_a = \{x \in \mathbb{R}^d : \exists J \subset [k] \text{ tal que } |J| \geq k/2 \text{ i, per a tot } j \in J,\ \|Z_j - x\| \leq \|Z_j - a\|\} \quad (10)$$

i definim l'estimador de la mitjana com a

$$\hat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} \operatorname{radi}(T_a),$$

on $\operatorname{radi}(T_a) = \sup_{x \in T_a} \|x - a\|$. Així, $\hat{\mu}_n$ es tria per a minimitzar, sobre tot $a \in \mathbb{R}^d$, el radi del conjunt T_a definit com el conjunt de punts $x \in \mathbb{R}^d$ per als quals $\|Z_j - x\| \leq \|Z_j - a\|$ per a la majoria dels blocs. Si hi ha diversos minimitzadors, se'n pot triar un qualsevol.

El conjunt T_a es pot veure com el conjunt de punts de $\mathbb{R}^d$ que estan almenys tan a prop del núvol de punts $\{Z_1, \ldots, Z_k\}$ com el punt a. L'estimador $\hat{\mu}_n$ s'obté minimitzant el radi de T_a.

Observem que sempre s'aconsegueix el mínim. Això es deu al fet que $\operatorname{radi}(T_a)$ és una funció contínua de a (ja que, per a cada a, T_a és la intersecció d'una unió finita de boles tancades, i els centres i radis de les boles tancades són contínues en a).

Es pot interpretar $\operatorname{argmin}_{a \in \mathbb{R}^d} \operatorname{radi}(T_a)$ com una altra noció multivariant de la mediana de $Z_1, \ldots, Z_k$. De fet, quan $d = 1$, és una elecció particular de la mediana i l'estimador coincideix amb l'estimador mediana de les mitjanes.

La fita de rendiment següent mostra que l'estimador té el rendiment subgaussià desitjat.

TEOREMA 8 (LUGOSI I MENDELSON [46]). *Sigui $\delta \in (0,1)$ i considerem l'estimador de la mitjana $\widehat{\mu}_n$ amb el paràmetre $k = \lceil 200 \log(2/\delta) \rceil$. Si $X_1, \ldots, X_n$ són vectors aleatoris i.i.d. a $\mathbb{R}^d$ amb mitjana $\mu \in \mathbb{R}^d$ i matriu de covariància Σ, aleshores per a tot n, amb probabilitat almenys $1-\delta$,*

$$\|\widehat{\mu}_n - \mu\| \le \max\left(960\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right).$$

Igual que la fita de rendiment de la proposició 7, el teorema 8 és «infinitodimensional» en el sentit que la fita no depèn explícitament de la dimensió d. De fet, el mateix estimador es pot definir per a vectors aleatoris que prenguin valors en espais de Hilbert i el teorema 8 continua sent vàlid mentre $\mathrm{Tr}(\Sigma) = \mathbb{E}\|X-\mu\|^2$ sigui finit.

El teorema 8 és desprèn del resultat següent.

TEOREMA 9. *Utilitzant la notació prèvia i fixant*

$$r = \max\left(960\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right),$$

amb probabilitat almenys $1-\delta$, per a qualsevol $a \in \mathbb{R}^d$ tal que $\|a-\mu\| \ge r$, es té $\|Z_j - a\| > \|Z_j - \mu\|$ per a més de $k/2$ índexs j. En altres paraules, $\|a-\mu\| \ge r$ implica que $a \notin T_\mu$.

El teorema 9 implica que per a una col·lecció «típica» $X_1, \ldots, X_n$, μ és més proper a la majoria dels Z_j en comparació amb qualsevol $a \in \mathbb{R}^d$ que estigui prou lluny de μ. Òbviament, per a una col·lecció arbitrària $x_1, \ldots, x_n \subset \mathbb{R}^d$, ni tan sols cal que existeixi aquest punt, i és sorprenent que per a una configuració típica i.i.d. μ satisfaci aquesta propietat.

El fet que el teorema 9 impliqui el teorema 8 és senzill. De fet, la definició de $\widehat{\mu}_n$ i el teorema 9 impliquen que, amb probabilitat almenys $1-\delta$, $\mathrm{radi}(T_{\widehat{\mu}_n}) \le \mathrm{radi}(T_\mu) \le r$. Com que, o bé $\mu \in T_{\widehat{\mu}_n}$ o bé $\widehat{\mu} \in T_\mu$, llavors segur que $\|\widehat{\mu}_n - \mu\| \le r$, tal com es demana.

Certament, les constants que apareixen al teorema 8 no són òptimes; es van obtenir amb l'objectiu de fer la demostració més transparent.

La demostració del teorema 9 es basa en la idea següent. La mitjana μ és el minimitzador de la funció $f(x) = \mathbb{E}\|X-x\|^2$. Un enfocament possible és utilitzar les dades disponibles per a endevinar, per a qualsevol parell $a, b \in \mathbb{R}^d$, si $f(a) < f(b)$. Una opció natural és fer servir un estimador mediana de mitjanes per tal de decidir quin dels dos és millor. El «torneig» és simplement una manera de comparar cada parell, tal com es descriu a continuació.[1]

1 Tal com expliquem a continuació, n'hi ha prou d'assegurar-se que la comparació és correcta entre μ i qualsevol punt que no estigui massa a prop de μ.

Per definir el torneig, recordem que $[n]$ està dividit en k blocs disjunts $B_1, \dots, B_k$ de mida $m = n/k$. Per a $a, b \in \mathbb{R}^d$, diem que *a derrota b* si

$$\frac{1}{m} \sum_{i \in B_j} (\|X_i - b\|^2 - \|X_i - a\|^2) > 0 \tag{11}$$

en més de $k/2$ blocs B_j. El principal lema tècnic és el següent.

LEMA 10. *Sigui $\delta \in (0,1)$, $k = \lceil 200 \log(2/\delta) \rceil$, i definim*

$$r = \max\left(960\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max} \log(2/\delta)}{n}}\right).$$

Amb una probabilitat almenys $1 - \delta$, μ derrota tots els $b \in \mathbb{R}^d$ de manera que $\|b - \mu\| \geq r$.

El resultat del lema 10 és raonable: si $\|b - \mu\|$ és prou gran, això es reflectirà en els «valors típics» de $(\|X_i - \mu\|)_{i=1}^n$ i $(\|X_i - b\|)_{i=1}^n$. La comparació dels valors mitjançant (11) assegura «estabilitat», i el fet que b estigui lluny de μ es mostra amb alta probabilitat. Subratllem que l'estimació de probabilitat ha de ser *uniforme* en b. Aquestes estimacions uniformes esdevindran un tema recurrent en el que resta d'article.

PROVA. Notem que

$$\|X_i - b\|^2 - \|X_i - \mu\|^2 = \|X_i - \mu + \mu - b\|^2 - \|X_i - \mu\|^2 = -2\langle X_i - \mu, b - \mu\rangle + \|b - \mu\|^2,$$

definim $\overline{X} = X - \mu$ i denotem $v = b - \mu$. Així, per a un b fix que satisfaci $\|b - \mu\| \geq r$, μ derrota b si

$$-\frac{2}{m} \sum_{i \in B_j} \langle \overline{X}_i, v\rangle + \|v\|^2 > 0$$

a la majoria dels blocs B_j.

Per tant, per a demostrar la nostra afirmació necessitem que, amb probabilitat almenys $1 - \delta$, per a cada $v \in \mathbb{R}^d$ amb $\|v\| \geq r$,

$$-\frac{2}{m} \sum_{i \in B_j} \langle \overline{X}_i, v\rangle + \|v\|^2 > 0 \tag{12}$$

en més de $k/2$ blocs B_j. Clarament, n'hi ha prou de demostrar que (12) es compleix quan $\|v\| = r$.

Considerem un $v \in \mathbb{R}^d$ fix amb $\|v\| = r$. Per la desigualtat de Txebixev, amb una probabilitat d'almenys $9/10$,

$$\left| \frac{1}{m} \sum_{i \in B_j} \langle \overline{X}_i, v\rangle \right| \leq \sqrt{10}\sqrt{\frac{\mathbb{E}\langle \overline{X}, v\rangle^2}{m}} \leq \sqrt{10}\|v\|\sqrt{\frac{\lambda_{\max}}{m}},$$

on recordem que $\lambda_{\max}$ és el valor propi més gran de la matriu de covariància Σ de X. Així, si

$$r = \|v\| \geq 4\sqrt{10}\sqrt{\frac{\lambda_{\max}}{m}}, \tag{13}$$

aleshores amb probabilitat almenys $9/10$,

$$-\frac{2}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle \geq \frac{-r^2}{2}. \tag{14}$$

Aplicant la desigualtat de Hoeffding ([24]), obtenim que (14) és certa per a un sol v amb probabilitat almenys $1-\exp(-k/50)$ en almenys $8/10$ dels blocs B_j.

Ara ho hem d'estendre a tots els vectors amb norma r a partir d'un vector fix v. Per tal de demostrar que (14) és certa simultàniament per a tots els $v \in r\cdot S^{d-1}$ en almenys $7/10$ dels blocs B_j, primer considerem un conjunt $V_1 \subset r\cdot S^{d-1}$ maximal i ϵ-separat respecte a la norma $L_2(X)$. En altres paraules, V_1 és un subconjunt de $r\cdot S^{d-1}$ de cardinalitat maximal tal que per a tots els $v_1, v_2 \in V_1$, $\|v_1-v_2\|_{L_2(X)} = \langle v_1-v_2, \Sigma(v_1-v_2)\rangle^{1/2} \geq \epsilon$. Podem estimar aquesta cardinalitat per la desigualtat «dual de Sudakov» (vegeu Ledoux i Talagrand [40] i també Vershynin [64] per a una versió amb la constant específica utilitzada aquí): la cardinalitat de V_1 està fitada per

$$\log(|V_1|/2) \leq \frac{1}{32}\left(\frac{\mathbb{E}[\langle G, \Sigma G\rangle^{1/2}]}{\epsilon/r}\right)^2,$$

on G és un vector normal estàndard a $\mathbb{R}^d$. Observem que per a qualsevol $a \in \mathbb{R}^d$, $\mathbb{E}_X\langle a, X\rangle^2 = \langle a, \Sigma a\rangle$ i, per tant,

$$\begin{aligned}\mathbb{E}[\langle G, \Sigma G\rangle^{1/2}] &= \mathbb{E}_G[(\mathbb{E}_X[\langle G, \overline{X}\rangle^2])^{1/2}] \leq (\mathbb{E}_X\mathbb{E}_G[\langle G, \overline{X}\rangle^2])^{1/2} = \\ &= (\mathbb{E}[\|\overline{X}\|^2])^{1/2} = \sqrt{\mathrm{Tr}(\Sigma)}.\end{aligned}$$

Per tant, fixant

$$\epsilon = 2r\left(\frac{1}{k}\mathrm{Tr}(\Sigma)\right)^{1/2}, \tag{15}$$

tenim $|V_1| \leq 2e^{k/100}$ i per la fita de la unió, amb probabilitat almenys $1-2e^{-k/100} \geq 1-\delta/2$, (14) és vàlida per a tots els $v \in V_1$ en almenys $8/10$ dels blocs B_j.

A continuació, comprovem que la propietat (12) es compleix simultàniament per a tots els x amb $\|x\| = r$ en almenys $7/10$ dels blocs B_j.

Per a cada $x \in r\cdot S^{d-1}$, sigui v_x l'element més proper a x en V_1 respecte a la norma $L_2(X)$. N'hi ha prou de demostrar que, amb probabilitat almenys $1-\exp(-k/200) \geq 1-\delta/2$,

$$\sup_{x\in r\cdot S^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\mathbb{1}_{\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle| \geq r^2/4\}} \leq \frac{1}{10}. \tag{16}$$

De fet, en aquest esdeveniment es dedueix que per a cada $x \in r \cdot S^{d-1}$, almenys en $7/10$ dels blocs B_j, tant

$$-\frac{2}{m}\sum_{i\in B_j}\langle \overline{X}_i, v_x\rangle \geq \frac{-r^2}{2} \quad \text{com} \quad 2\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, x\rangle - \frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, v_x\rangle\right| < \frac{r^2}{2}$$

se satisfan i, per tant, en aquests blocs, $-\frac{2}{m}\sum_{i\in B_j}\langle \overline{X}_i, x\rangle + r^2 > 0$, tal com volíem.

Queda per demostrar (16). Observeu que

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle| \geq r^2/4\}} \leq \frac{4}{r^2}\frac{1}{k}\sum_{j=1}^{k}\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle\right|.$$

Com que $\|x - v_x\|_{L_2(X)} = \sqrt{\mathbb{E}\langle \overline{X}, x - v_x\rangle^2} \leq \epsilon$, es dedueix que, per a cada j,

$$\mathbb{E}\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle\right| \leq \sqrt{\frac{\mathbb{E}[\langle \overline{X}, x-v_x\rangle^2]}{m}} \leq \frac{\epsilon}{\sqrt{m}},$$

i, per tant,

$$\mathbb{E}\sup_{x\in r\cdot S^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\mathbb{1}_{\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle| \geq r^2/4\}} \leq$$

$$\leq \underbrace{\frac{4}{r^2}\mathbb{E}\sup_{x\in r\cdot S^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\left(\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle\right| - \mathbb{E}\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle\right|\right)}_{(A)} + \underbrace{\frac{4\epsilon}{r^2\sqrt{m}}}_{(B)}.$$

Per fitar (B), notem que, usant (15),

$$\frac{4\epsilon}{r^2\sqrt{m}} = 8\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}\cdot\frac{1}{r} \leq \frac{1}{60}$$

sempre que

$$r \geq 480\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}.$$

Podem fitar (A) mitjançant tècniques estàndard de processos empírics com ara la simetrització, la contracció per a mitjanes de Rademacher i la dessimetrització. En efecte, si $\sigma_1, \ldots, \sigma_n$ són variables aleatòries de Rademacher

independents (és a dir, $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$), independentment de tots els X_i, aleshores

$$(A) \le \frac{8}{r^2}\mathbb{E} \sup_{x \in r \cdot S^{d-1}} \frac{1}{k}\sum_{j=1}^{k} \sigma_j \left| \frac{1}{m} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle \right| \le$$

(per una desigualtat de simetrització estàndard;

vegeu, per exemple, [62, Lemma 2.3.6])

$$\le \frac{8}{r^2}\mathbb{E} \sup_{x \in r \cdot S^{d-1}} \left| \frac{1}{k}\sum_{j=1}^{k} \sigma_j \frac{1}{m} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle \right| \le$$

(per un lema de contracció per a mitjanes de Rademacher; vegeu [40])

$$\le \frac{16}{r^2}\mathbb{E} \sup_{x \in r \cdot S^{d-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \langle \overline{X}_i, x - v_x \rangle \right| \le$$

(vegeu una altra vegada [62, Lemma 2.3.6])

$$\le \frac{32}{r}\mathbb{E} \sup_{\{t : \|t\| \le 1\}} \left| \frac{1}{n}\sum_{i=1}^{n} \langle \overline{X}_i, t \rangle \right| \le$$

(observem que $\|x - v_x\| \le 2r$)

$$\le \frac{32}{r} \cdot \frac{\sqrt{\mathbb{E}\|\overline{X}\|^2}}{\sqrt{n}} = \frac{32}{r}\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2} \le \frac{1}{30}$$

sempre que $r \ge 960\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}$.

Així, per a

$$Y = \sup_{x \in r \cdot S^{d-1}} \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{|m^{-1}\sum_{i \in B_j}\langle \overline{X}_i, x - v_x \rangle| \ge r^2/4\}},$$

hem demostrat que $\mathbb{E}Y \le 1/60 + 1/30 = 1/20$. Finalment, per a obtenir (16), n'hi ha prou de demostrar que $\mathbb{P}\{Y > \mathbb{E}Y + 1/20\} \le e^{-k/200}$, que es desprèn de la desigualtat de diferències fitades (vegeu, per exemple, [6, teorema 6.2]). □

Demostració del teorema 9 El teorema 9 es dedueix fàcilment del lema 10. Fixem un bloc B_j i recordem que $Z_j = \frac{1}{m}\sum_{i \in B_j} X_i$. Siguin $a, b \in \mathbb{R}^d$. Aleshores,

$$\frac{1}{m}\sum_{i \in B_j}(\|X_i - a\|^2 - \|X_i - b\|^2) = \frac{1}{m}\sum_{i \in B_j}(\|X_i - b - (a - b)\|^2 - \|X_i - b\|^2) =$$

$$= -\frac{2}{m}\sum_{i \in B_j}\langle X_i - b, a - b \rangle + \|a - b\|^2 = (*).$$

Observem que $-\frac{2}{m}\sum_{i\in B_j}\langle X_i-b, a-b\rangle = -2\langle \frac{1}{m}\sum_{i\in B_j} X_i - b, a-b\rangle = -2\langle Z_j - b, a-b\rangle$, i, per tant,

$$\begin{aligned}(*) &= -2\langle Z_j - b, a-b\rangle + \|a-b\|^2 = \\ &= -2\langle Z_j - b, a-b\rangle + \|a-b\|^2 + \|Z_j - b\|^2 - \|Z_j - b\|^2 = \\ &= \|Z_j - b - (a-b)\|^2 - \|Z_j - b\|^2 = \|Z_j - a\|^2 - \|Z_j - b\|^2.\end{aligned}$$

En conseqüència, $(*) > 0$ (és a dir, b derrota a al bloc B_j) si i només si $\|Z_j - a\| > \|Z_j - b\|$.

Recordem que el lema 10 estableix que, amb probabilitat almenys $1-\delta$, si $\|a-\mu\| \geq r$, llavors en més de $k/2$ blocs B_j, $\frac{1}{m}\sum_{i\in B_j}(\|X_i - a\|^2 - \|X_i - \mu\|^2) > 0$, que, segons l'argument anterior, és el mateix que dir que almenys per a $k/2$ índexs j, es compleix que $\|Z_j - a\| > \|Z_j - \mu\|$.

Reflexionant-hi, queda clar que les idees utilitzades en la demostració del teorema 8 són força generals. De fet, són al cor del *mètode de la bola petita* introduït per Mendelson [48] (vegeu també [49] per a mètodes de caire similar). El mètode de la bola petita s'aplica en situacions molt més generals que el teorema 8. Per tal d'explicar com se'n pot estendre l'argument, tornem a esbossar els tres passos que ens han permès comparar cada b i μ:

(1) Per a qualsevol b *fixat* de $\mathbb{R}^d$, obtenim una fita que es compleix amb una alta probabilitat.

(2) Aleshores, gràcies a l'estimació amb alta probabilitat de (1), podem invocar la fita de la unió i controlar una col·lecció gran (però encara finita) de punts.

    Tenim total llibertat per a triar aquesta col·lecció, i la seleccionem com una xarxa de radi ϵ al conjunt en qüestió.

(3) La part crucial de l'argument és passar del control que tenim en cada punt de la xarxa al control uniforme desitjat a tota la classe; concretament, mostrem que si un «centre», és a dir, un element de la xarxa, té un bon comportament,[2] llavors el mateix és cert per a qualsevol punt prou proper al centre. Amb aquesta finalitat, mostrem que les «oscil·lacions aleatòries» no destrueixen el bon comportament d'un centre en gaires blocs.

3.5 Consideracions computacionals

Un tema important que hem passat de llarg fins a aquest punt és la viabilitat computacional dels estimadors de mitjanes. Tot i que la mitjana empírica és trivial de calcular, molts dels estimadors més sofisticats que hem discutit en aquest article estan lluny de ser-ho. En particular, un requisit bàsic perquè qualsevol estimador multivariant de la mitjana sigui útil a la pràctica és que es pugui calcular en temps polinomial (és a dir, en un temps que sigui un polinomi de la mida de la mostra n i de la dimensió d). Com ja hem assenyalat,

2 A la demostració del teorema 8, «bon comportament» significa que (12) és vàlida per a la majoria dels blocs.

alguns dels estimadors descrits anteriorment entren en aquesta categoria. Per exemple, l'estimador mediana geomètrica de les mitjanes o l'estimador de Catoni-Giulini són calculables de manera eficient en aquest sentit. No obstant això, aquests estimadors estan lluny de ser subgaussians. L'estimador de torneig de medianes de mitjanes és subgaussià, però el seu càlcul suposa un repte gens trivial. De fet, tal com es defineix l'estimador, és probable que sigui computacionalment intractable (és a dir, NP-complex). No obstant això, en un article recent, Hopkins ([25]) defineix una relaxació semidefinida enginyosa de l'estimador de torneig de medianes de mitjanes que es pot calcular en temps $O(nd + d\log(1/\delta)^c)$ per a una constant independent de la dimensió i, al mateix temps, garanteix la condició subgaussiana desitjada sota l'únic supòsit que existeixi la matriu de covariància. Aquest és el primer estimador de la mitjana multivariant subgaussià calculable de manera eficient. Encara més recentment, Cherapanamjeri, Flammarion i Bartlett ([13]) van millorar el temps d'execució a $O(nd + d\log(1/\delta)^2 + \log(1/\delta)^4)$ combinant les idees de Hopkins amb una optimització astuta de gradient descendent. Basant-se en idees de «reponderació espectral» de Cheng, Diakonikolas i Ge [12], Depersin i Lecué [16] i Lei, Luh, Venkat i Zhang [42] milloren encara més el temps d'execució. Hopkins, Li i Zhang ([27]) mostren com la reponderació espectral és essencialment equivalent a la noció de mediana introduïda anteriorment. Remetem a aquests articles per a una revisió exhaustiva de la literatura sobre els aspectes computacionals de l'estimació de la mitjana robusta, que està creixent ràpidament.

A la comunitat de la informàtica teòrica hi ha hagut recentment un augment important de resultats que aborden el problema de l'estimació de la mitjana *robusta* i computacionalment eficient. En aquest context, un estimador es defineix com a robust si funciona bé en presència d'una petita fracció constant de valors atípics (possiblement antagonistes). S'han proposat diversos models; vegeu Charikar, Steinhardt i Valiant [11], Diakonikolas, Kamath, Kane, Li, Moitra i Stewart [18, 19, 20], Diakonikolas, Kane i Stewart [21], Diakonikolas, Kong i Stewart [22], Hopkins i Li [26], Klivans, Kothari i Meka [35], Kothari, Steinhardt i Steurer [37], Lai, Rao i Vempala [38], Loh i Tan [44], per a tenir una mostra d'aquest important gruix de literatura en plena expansió. La revisió d'aquesta àrea, però, va més enllà de l'abast d'aquest article.

Referències

[1] Alon, N.; Matias, Y.; Szegedy, M. «The space complexity of approximating the frequency moments». Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996). *J. Comput. System Sci.*, 58 (1), part 2 (1999), 137–147.

[2] Aloupis, G. «Geometric measures of data depth». A: *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*. Provi-

dence, RI: Amer. Math. Soc., 2006, 147–158. (DIMACS Ser. Discrete Math. Theoret. Comput. Sci.; 72)

[3] Bickel, P. J. «On some robust estimates of location». *Ann. Math. Statist.*, 36 (1965), 847–858.

[4] Blackwell, D. «Conditional expectation and unbiased sequential estimation». *Ann. Math. Statistics*, 18 (1947), 105–110.

[5] Blumer, A.; Ehrenfeucht, A.; Haussler, D.; Warmuth, M. K. «Learnability and the Vapnik-Chervonenkis dimension». *J. Assoc. Comput. Mach.*, 36 (4) (1989), 929–965.

[6] Boucheron, S.; Lugosi, G.; Massart, P. *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Prefaci de Michel Ledoux. Oxford: Oxford University Press, 2013.

[7] Bubeck, S.; Cesa-Bianchi, N.; Lugosi, G. «Bandits with heavy tail». *IEEE Trans. Inform. Theory*, 59 (11) (2013), 7711–7717.

[8] Catoni, O. «Challenging the empirical mean and empirical variance: a deviation study». *Ann. Inst. Henri Poincaré Probab. Stat.*, 48 (4) (2012), 1148–1185.

[9] Catoni, O.; Giulini, I. «Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression». Preprint (2017). [Disponible en línia a: `arXiv:1712.02747`]

[10] Catoni, O.; Giulini, I. «Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector». Preprint (2018). [Disponible en línia a: `arXiv:1802.04308`]

[11] Charikar, M.; Steinhardt, J.; Valiant, G. «Learning from untrusted data». A: *STOC'17—Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. Nova York: ACM, 2017, 47–60.

[12] Cheng, Y.; Diakonikolas, I.; Ge, R. «High-dimensional robust mean estimation in nearly-linear time». A: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Filadèlfia, PA: SIAM, 2019, 2755–2771.

[13] Cherapanamjeri, Y.; Flammarion, N.; Bartlett, P. L. «Fast mean estimation with sub-Gaussian rates». Preprint (2019). [Disponible en línia a: `arXiv:1902.01998`]

[14] Cirel'son, B. S.; Ibragimov, I. A.; Sudakov, V. N. «Norms of Gaussian sample functions». A: *Proceedings of the Third Japan-USSR Symposium on Probability Theory*. Berlín: Springer, 1976, 20–41. (Lecture Notes in Math.; 550)

[15] Cohen, M. B.; Lee, Y. T.; Miller, G.; Pachocki, J.; Sidford, A. «Geometric median in nearly linear time». A: *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. Nova York: ACM, 2016, 9–21.

[16] Depersin, J.; Lecué, G. «Robust subgaussian estimation of a mean vector in nearly linear time». Preprint (2019). [Disponible en línia a: `arXiv:1906.03058`]

[17] Devroye, L.; Lerasle, M.; Lugosi, G.; Oliveira, R. I. «Sub-Gaussian mean estimators». *Ann. Statist.*, 44 (6) (2016), 2695–2725.

[18] Diakonikolas, I.; Kamath, G.; Kane, D. M.; Li, J.; Moitra, A.; Stewart, A. «Robust estimators in high dimensions without the computational intractability». A: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. Los Alamitos, CA: IEEE Computer Soc., 2016, 655–664.

[19] Diakonikolas, I.; Kamath, G.; Kane, D. M.; Li, J.; Moitra, A.; Stewart, A. «Being robust (in high dimensions) can be practical». A: *ICML'17: Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Sydney, NSW, Austràlia: JMLR.org, 2017, 999–1008.

[20] Diakonikolas, I.; Kamath, G.; Kane, D. M.; Li, J.; Moitra, A.; Stewart, A. «Robustly learning a Gaussian: getting optimal error, efficiently». A: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. Filadèlfia, PA: SIAM, 2018, 2683–2702.

[21] Diakonikolas, I.; Kane, D. M.; Stewart, A. «Efficient robust proper learning of log-concave distributions». Preprint (2016). [Disponible en línia a: `arXiv:1606.03077`]

[22] Diakonikolas, I.; Kong, W.; Stewart, A. «Efficient algorithms and lower bounds for robust linear regression». A: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Filadèlfia, PA: SIAM, 2019, 2745–2754.

[23] Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J.; Stahel, W. A. *Robust Statistics. The Approach Based on Influence Functions*. Nova York: John Wiley & Sons, Inc., 1986. (Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics)

[24] Hoeffding, W. «Probability inequalities for sums of bounded random variables». *J. Amer. Statist. Assoc.*, 58 (1963), 13–30.

[25] Hopkins, S. B. «Mean estimation with sub-Gaussian rates in polynomial time». *Ann. Statist.*, 48 (2) (2020), 1193–1213.

[26] Hopkins, S. B.; Li, J. «Mixture models, robustness, and sum of squares proofs». A: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. Nova York: ACM, 2018, 1021–1034.

[27] Hopkins, S. B.; Li, J.; Zhang, F. «Robust and heavy-tailed mean estimation made simple, via regret minimization». Preprint (2020). [Disponible en línia a: `arXiv:2007.15839`]

[28] Hsu, D. «Robust statistics» (2010). [No publicat]

[29] Hsu, D.; Sabato, S. «Loss minimization and parameter estimation with heavy tails». *J. Mach. Learn. Res.*, 17, treball núm. 18 (2016), 40 p.

[30] Huber, M. «An optimal (ϵ, δ)-randomized approximation scheme for the mean of random variables with bounded relative variance». *Random Structures Algorithms*, 55 (2) (2019), 356–370.

[31] Huber, P. J. «Robust estimation of a location parameter». *Ann. Math. Statist.*, 35 (1964), 73–101.

[32] Huber, P. J.; Ronchetti, E. M. *Robust Statistics*. 2a ed. Hoboken, NJ: John Wiley & Sons, Inc., 2009. (Wiley Series in Probability and Statistics)

[33] Jerrum, M. R.; Valiant, L. G.; Vazirani, V. V. «Random generation of combinatorial structures from a uniform distribution». *Theoret. Comput. Sci.*, 43 (2-3) (1986), 169–188.

[34] Joly, E.; Lugosi, G.; Oliveira, R. I. «On the estimation of the mean of a random vector». *Electron. J. Stat.*, 11 (1) (2017), 440–451.

[35] Klivans, A.; Kothari, P. K.; Meka, R. «Efficient algorithms for outlier-robust regression». A: *Proceedings of Machine Learning Research*. Vol. 75. PMLR, 2018, 1420–1430.

[36] Kolmogorov, A. N. «Unbiased estimates». *Izvestiya Akad. Nauk SSSR. Ser. Mat.*, 14 (1950), 303–326. [En rus]

[37] Kothari, P. K.; Steinhardt, J.; Steurer, D. «Robust moment estimation and improved clustering via sum of squares». A: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. Nova York: ACM, 2018, 1035–1046.

[38] Lai, K. A.; Rao, A. B.; Vempala, S. «Agnostic estimation of mean and covariance». A: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. Los Alamitos, CA: IEEE Computer Soc., 2016, 665–674.

[39] Ledoux, M. *The Concentration of Measure Phenomenon*. Providence, RI: American Mathematical Society, 2001. (Mathematical Surveys and Monographs; 89)

[40] Ledoux, M.; Talagrand, M. *Probability in Banach Spaces. Isoperimetry and Processes*. Berlín: Springer-Verlag, 1991. (Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3); 23]

[41] Lee, J. C. H.; Valiant, P. «Optimal sub-Gaussian mean estimation in $\mathbb{R}$». A: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021*. Los Alamitos, CA: IEEE Computer Soc., 2022, 672–683.

[42] Lei, Z.; Luh, K.; Venkat, P.; Zhang, F. «A fast spectral algorithm for mean estimation with sub-Gaussian rates». A: *Proceedings of Machine Learning Research*. Vol. 125, PMLR, 2020, 2598–2612.

[43] Lerasle, M.; Oliveira, R. I. «Robust empirical mean estimators». Preprint (2011). [Disponible en línia a: `arXiv:1112.3914`]

[44] Loh, P.-L.; Tan, X. L. «High-dimensional robust precision matrix estimation: cellwise corruption under ϵ-contamination». *Electron. J. Stat.*, 12 (1) (2018), 1429–1467.

[45] Lugosi, G.; Mendelson, S. «Mean estimation and regression under heavy-tailed distributions: a survey». *Found. Comput. Math.*, 19 (5) (2019), 1145–1190.

[46] Lugosi, G.; Mendelson, S. «Sub-Gaussian estimators of the mean of a random vector». *Ann. Statist.*, 47 (2) (2019), 783–794.

[47] Lugosi, G.; Mendelson, S. «Robust multivariate mean estimation: the optimality of trimmed mean». *Ann. Statist.*, 49 (1) (2021), 393–410.

[48] Mendelson, S. «Learning without concentration». *J. ACM*, 62 (3) (2015), Art. 21, 25 p.

[49] Mendelson, S. «Learning without concentration for general loss functions». *Probab. Theory Related Fields*, 171 (1-2) (2018), 459–502.

[50] Minsker, S. «Geometric median and robust estimation in Banach spaces». *Bernoulli*, 21 (4) (2015), 2308–2335.

[51] Minsker, S.; Ndaoud, M. «Robust and efficient mean estimation: an approach based on the properties of self-normalized sums». *Electron. J. Stat.*, 15 (2) (2021), 6036–6070.

[52] Minsker, S.; Strawn, N. «Distributed statistical estimation and rates of convergence in normal approximation». Preprint (2017). [Disponible en línia a: `arXiv:1704.02658`]

[53] Nemirovsky, A. S.; Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience Publication. Nova York: John Wiley & Sons, Inc., 1983. (Wiley-Interscience Series in Discrete Mathematics). [Traduït del rus i amb prefaci d'E. R. Dawson]

[54] Oliveira, R. I.; Orenstein, P. «The sub-gaussian property of trimmed means estimators». Informe tècnic, IMPA (2019).

[55] Radhakrishna Rao, C. «Information and the accuracy attainable in the estimation of statistical parameters». *Bull. Calcutta Math. Soc.*, 37 (1945), 81–91.

[56] Shevtsova, I. G. «On absolute constants in inequalities of Berry-Esseen type». *Dokl. Akad. Nauk*, 456 (6) (2014), 650–654. [En rus; traducció a: *Dokl. Math.*, 89 (3) (2014), 378–381]

[57] Small, C. G. «A survey of multidimensional medians». *Int. Stat. Rev.*, 58 (3) (1990), 263–277.

[58] Stigler, S. M. «The asymptotic distribution of the trimmed mean». *Ann. Statist.*, 1 (1973), 472–477.

[59] Tukey, J. W. «Mathematics and the picturing of data». A: *Proceedings of the International Congress of Mathematicians*. Vol. 2. Montreal, Que.: Canad. Math. Congress, 1975, 523–531.

[60] Tukey, J. W.; McLaughlin, D. H. «Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. I». *Sankhyā Ser. A*, 25 (1963), 331–352.

[61] Valiant, L. G. «A theory of the learnable». *Communications of the ACM*, 27 (11) (1984), 1134–1142.

[62] van der Vaart, A. W.; Wellner, J. A. *Weak Convergence and Empirical Processes. With Applications to Statistics*. Nova York: Springer-Verlag, 1996. (Springer Series in Statistics)

[63] Vapnik, V. N.; Chervonenkis, A. Ya. *Theory of Pattern Recognition*. Moscou: Statistical problems of learning Izdat. «Nauka», 1974. [En rus]

[64] Vershynin, R. «Lectures in Geometric Functional Analysis» (2009). [No publicat]

Departament d'Economia i Empresa, Universitat Pompeu Fabra
ICREA, Pg. Lluís Companys 23, 08010 Barcelona
Barcelona Graduate School of Economics
gabor.lugosi@upf.edu