Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector

Olivier Catoni
CREST – CNRS UMR 9194
Université Paris Saclay
olivier.catoni@ensae.fr

Ilaria Giulini
Laboratoire de Probabilités
et Modèles Aléatoires
Université Paris Diderot
giulini@math.univ-paris-diderot.fr

Abstract

In this paper, we present a new estimator of the mean of a random vector, computed by applying some threshold function to the norm. Non asymptotic dimension-free almost sub-Gaussian bounds are proved under weak moment assumptions, using PAC-Bayesian inequalities.

1 Introduction

Estimating the mean of a random vector under weak tail assumptions has attracted a lot of attention recently. A number of properties have spurred the interest for these new results, where the empirical mean is replaced by a more robust estimator. One aspect is that it is possible to obtain an estimator with a sub-Gaussian tail while assuming much weaker assumptions on the data, up to the fact of assuming only the existence of a finite covariance matrix. Another appealing feature is that it is possible to obtain dimension-free non asymptotic bounds that remain valid in a separable Hilbert space. Some important references are Catoni [2012] in the one dimensional case and Minsker [2015] and Lugosi and Mendelson [2017] in the multidimensional case. Building on the breakthrough of Minsker [2015], that uses a multidimensional generalization of the median of means estimator, Joly et al. [2017] and Lugosi and Mendelson [2017] propose successive improvements of the median of means approach to get an estimator with a genuine sub-Gaussian dimension-free tail bound, while still requiring only the existence of the covariance matrix. In the mean time, the M-estimator approach of Catoni [2012] has also been generalized to multidimensional settings through the use of matrix inequalities in Minsker [2016] and Minsker and Wei [2017].

Here we follow a different route, based on a multidimensional extension of Catoni [2012] using PAC-Bayesian bounds. Our new estimator is a simple modification of the empirical mean, where some threshold is applied to the norm of the sample vectors. Therefore, it is straightforward to compute, and this is a strong point of our approach, compared to others. Note also that we make here some compromise on the sharpness of the estimation error bound, in order to simplify the definition and computation of the estimator. This compromise consists in the presence of second order terms, while the first order terms can be made as close as desired to a true sub-Gaussian bound with exact constants, as stated in Lugosi and Mendelson [2017, eq. (1.1)]. With a more involved estimator, a true sub-Gaussian bound without second order terms is possible and will be described in a separate publication.

2 Thresholding the norm

Consider $X \in \mathbb{R}^d$, a random vector, and $(X_1, \dots, X_n)$ a sample made of n independent copies of X. The question is to estimate $\mathbb{E}(X)$ from the sample, under the assumption that $\mathbb{E}\big(\|X\|^p\big) < \infty$, for some $p \geq 2$.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Consider the threshold function $\psi(t) = \min\{t, 1\},\ t \in \mathbb{R}_+$, and for some positive real parameter λ to be chosen later, introduce the thresholded sample

$$Y_i = \frac{\psi\big(\lambda\|X_i\|\big)}{\lambda\|X_i\|} X_i.$$

Our estimator of $m = \mathbb{E}(X)$ will simply be the thresholded empirical mean $\widehat{m} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i$.

Proposition 2.1 *Introduce the increasing functions*

$$g_1(t) = \frac{1}{t}\Big(\exp(t) - 1\Big) \text{ and } g_2(t) = \frac{2}{t^2}\Big(\exp(t) - 1 - t\Big), \qquad t \in \mathbb{R},$$

that are defined by continuity at $t = 0$ and are such that $g_1(0) = g_2(0) = 1$. Assume that $\mathbb{E}\big(\|X\|^2\big) < \infty$ and that we know v such that

$$\sup_{\theta \in \mathbb{S}_d} \mathbb{E}\big(\langle \theta, X - m\rangle^2\big) \le v < \infty,$$

where $\mathbb{S}_d = \big\{\theta \in \mathbb{R}^d, \|\theta\| = 1\big\}$ is the unit sphere of $\mathbb{R}^d$. For some positive real parameter μ, put

$$\lambda = \mu^{-1}\sqrt{\frac{2\log(\delta^{-1})}{avn}}, \qquad T = \max\big\{\mathbb{E}\big(\|X - m\|^2\big), v\big\}, \qquad a = g_2(2\mu) \ge 1,$$
$$b = \exp(2\mu) g_1\left(\frac{\sqrt{2a}\mu^2}{\sqrt{\log(\delta^{-1})}}\right) \ge 1, \qquad \beta = \sqrt{\frac{2bT\log(\delta^{-1})}{av}} \ge \sqrt{\frac{2b\log(\delta^{-1})}{a}}.$$

With probability at least $1 - \delta$,

$$\|\widehat{m} - m\| \le \sqrt{\frac{2av\log(\delta^{-1})}{n}} + \sqrt{\frac{bT}{n}} + \inf_{p \ge 1} \frac{C_p}{n^{p/2}} + \inf_{p \ge 2} \frac{C'_p}{n^{p/2}},$$

where

$$C_p = \frac{1}{p+1}\left(\frac{p}{(p+1)\mu}\right)^p \left(\frac{2\log(\delta^{-1})}{av}\right)^{p/2} \sup_{\theta \in \mathbb{S}_d} \mathbb{E}\big(\|X\|^p \langle \theta, X - m\rangle_-\big), \quad \text{and}$$
$$C'_p = \frac{1}{p+1}\left(\frac{p}{(p+1)\mu}\right)^p \left(\frac{2\log(\delta^{-1})}{av}\right)^{p/2} \mathbb{E}\big(\|X\|^p\big)\|m\| \left(1 + \sqrt{\frac{\log(\delta^{-1})}{2avn}}\Big(a + \frac{b}{\beta}\Big)\|m\|\right).$$

Remarks Note that in case $\mathbb{E}\big(\|X\|^2\big) < \infty$ but $\mathbb{E}\big(\|X\|^p\big) = \infty$ for $p > 2$, we can use the bound

$$\frac{C_1}{\sqrt{n}} + \frac{C'_2}{n} \le \frac{1}{2\mu}\sqrt{\frac{\log(\delta^{-1})(T + \|m\|^2)}{2an}} + \frac{8\log(\delta^{-1})}{27\mu^2 avn}\mathbb{E}\big(\|X\|^2\big)\|m\|$$
$$\times \left(1 + \sqrt{\frac{\log(\delta^{-1})}{2avn}}\Big(a + \frac{b}{\beta}\Big)\|m\|\right) = \mathcal{O}\left(\frac{1}{2\mu}\sqrt{\frac{\log(\delta^{-1})(T + \|m\|^2)}{2an}}\right).$$

Note also that if we take $\mu = 1/4$ and assume that $\delta \le \exp(-1)$, then $a \le 1.2$ and $b \le 4$. If moreover $\mathbb{E}\big(\|X\|^{p+1}\big) < \infty$, for some $p > 1$, we obtain with probability at least $1 - \delta$ that

$$\|\widehat{m} - m\| \le \sqrt{\frac{2.4\, v\log(\delta^{-1})}{n}} + \sqrt{\frac{4T}{n}} + \frac{C_p}{n^{p/2}} + \frac{C'_{p+1}}{n^{(p+1)/2}},$$

meaning that the tail distribution of $\|\widehat{m} - m\|$ has a sub-Gaussian behavior, up to second order terms. Remark that by taking μ small, we can make a and b as close as desired to 1, at the expense of the values of C_p and C'_p.

Proof The rest of the paper is devoted to the proof of Proposition 2.1.

An elementary computation shows that the threshold function ψ satisfies

$$0 \leq 1 - \frac{\psi(t)}{t} \leq \inf_{p \geq 1} \frac{t^p}{p+1} \left(\frac{p}{p+1} \right)^p, \qquad t \in \mathbb{R}_+, \tag{1}$$

where non integer values of the exponent p are allowed. Let $Y = \dfrac{\psi\big(\lambda \|X\|\big)}{\lambda \|X\|} X$ and $\widetilde{m} = \mathbb{E}(Y)$. We can decompose the estimation error in direction θ into

$$\langle \theta, \widehat{m} - m \rangle = \langle \theta, \widetilde{m} - m \rangle + \frac{1}{n} \sum_{i=1}^n \langle \theta, Y_i - \widetilde{m} \rangle, \qquad \theta \in \mathbb{R}^d. \tag{2}$$

Introduce $\alpha = \dfrac{\psi\big(\lambda \|X\|\big)}{\lambda \|X\|}$ and let us deal with the first term first. As $0 \leq 1 - \alpha \leq \dfrac{\lambda^p \|X\|^p}{p+1} \left(\dfrac{p}{p+1} \right)^p$

$$\begin{aligned}
&\langle \theta, \widetilde{m} - m \rangle = \mathbb{E}\big[(\alpha - 1) \langle \theta, X \rangle\big] = \mathbb{E}\big[(\alpha - 1) \langle \theta, X - m \rangle\big] + \mathbb{E}(\alpha - 1) \langle \theta, m \rangle \\
&\leq \inf_{p \geq 1} \frac{\lambda^p}{(p+1)} \left(\frac{p}{p+1} \right)^p \mathbb{E}\Big(\|X\|^p \langle \theta, X - m \rangle_- \Big) + \inf_{p \geq 2} \frac{\lambda^p}{(p+1)} \left(\frac{p}{p+1} \right)^p \mathbb{E}\big(\|X\|^p \big) \langle \theta, m \rangle_-,
\end{aligned}$$

where $r_- = \max\{0, -r\}$ is the negative part of integer r.

Let us now look at the second term of the decomposition (2). To gain uniformity in θ, we will use a PAC-Bayesian inequality and the family of normal distributions $\rho_\theta = \mathcal{N}\big(\theta, \beta^{-1} I_d\big)$, bearing on the parameter $\theta \in \mathbb{R}^d$, where $I_d \in \mathbb{R}^{d \times d}$ is the identity matrix of size $d \times d$.

We will use the following PAC-Bayesian inequality without recalling its proof, that is a simple consequence of Catoni [2004, eq. (5.2.1) page 159]:

Lemma 2.2 *For any bounded measurable function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, for any probability measure $\pi \in \mathcal{M}_+^1\big(\mathbb{R}^d\big)$, for any $\delta \in]0,1[$, with probability at least $1 - \delta$, for any probability measure $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$,*

$$\frac{1}{n} \sum_{i=1}^n \int f\big(\theta, X_i\big) \, \mathrm{d}\rho(\theta) \leq \int \log\Big[\mathbb{E}\Big(\exp\big(f(\theta, X) \big) \Big) \Big] \, \mathrm{d}\rho(\theta) + \frac{\mathcal{K}(\rho, \pi) + \log(\delta^{-1})}{n},$$

where $\mathcal{K}$ is the Kullback-Liebler divergence $\mathcal{K}(\rho, \pi) = \begin{cases} \int \log\big(\rho/\pi\big) \, \mathrm{d}\rho, & \text{when } \rho \ll \pi, \\ +\infty, & \text{otherwise.} \end{cases}$

Remarking that $\dfrac{1}{n} \sum_{i=1}^n \langle \theta, Y_i - \widetilde{m} \rangle = \dfrac{1}{n} \sum_{i=1}^n \int \langle \theta', Y_i - \widetilde{m} \rangle \, \mathrm{d}\rho_\theta(\theta')$, using $\pi = \rho_0$, and taking into account the fact that $\mathcal{K}(\rho_\theta, \rho_0) = \beta \|\theta\|^2 / 2$, we obtain as a consequence of the previous lemma that with probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$,

$$\frac{1}{n} \sum_{i=1}^n \langle \theta, Y_i - \widetilde{m} \rangle \leq \frac{1}{\mu\lambda} \int \log\Big(\mathbb{E} \exp\Big(\mu\lambda \langle \theta', Y - \widetilde{m} \rangle \Big) \Big) \, \mathrm{d}\rho_\theta(\theta') + \frac{\beta}{2n\mu\lambda} + \frac{\log(\delta^{-1})}{n\mu\lambda}.$$

In our setting f is not bounded in θ, but the required extension is valid as explained in Catoni [2004]. Since the logarithm is concave,

$$\begin{aligned}
\int \log\Big(\mathbb{E} \exp\Big(\mu\lambda \langle \theta', Y - \widetilde{m} \rangle \Big) \Big) \, \mathrm{d}\rho_\theta(\theta') &\leq \log\left[\mathbb{E}\Big(\int \exp\Big(\mu\lambda \langle \theta', Y - \widetilde{m} \rangle \Big) \, \mathrm{d}\rho_\theta(\theta') \Big) \right] \\
&= \log\left[\mathbb{E}\Big(\exp\Big(\mu\lambda \langle \theta, Y - \widetilde{m} \rangle + \frac{\mu^2 \lambda^2}{2\beta} \|Y - \widetilde{m}\|^2 \Big) \Big) \right],
\end{aligned}$$

where we have used the explicit expression of the Laplace transform of a Gaussian distribution.

To go further, reminding as a source of inspiration the proof of Bennett's inequality, let us introduce the increasing functions g_1 and g_2 defined in Proposition 2.1. These functions will be used to bound the exponential function by polynomials. More precisely, we will exploit the fact that when $t \leq b$, $\exp(t) \leq 1 + t + g_2(b)t^2/2$ and $\exp(t) \leq 1 + g_1(b)t$. From this, it results that if $t \leq b$ and $u \leq c$,

$$\begin{aligned}\exp(t+u) \leq \exp(t)\big(1 + g_1(c)u\big) &\leq \exp(t) + g_1(c)\exp(b)u \\ &\leq 1 + t + g_2(b)t^2/2 + g_1(c)\exp(b)u.\end{aligned}$$

Legitimate values for b and c will be deduced from the remark that $\lambda\|Y\| \leq 1$, implying $\lambda\|\widetilde{m}\| \leq 1$. Namely, in our context, we will use $b = 2\mu$ and $c = 2\mu^2/\beta$.

These arguments put together lead to the inequality

$$\begin{aligned}&\mathbb{E}\left(\exp\left(\mu\lambda\langle\theta, Y - \widetilde{m}\rangle + \frac{\mu^2\lambda^2}{2\beta}\big\|Y - \widetilde{m}\big\|^2\right)\right) \\ &\qquad\leq 1 + g_2(2\mu)\frac{\mu^2\lambda^2}{2}\mathbb{E}\big(\langle\theta, Y - \widetilde{m}\rangle^2\big) + \exp(2\mu)g_1\left(\frac{2\mu^2}{\beta}\right)\frac{\mu^2\lambda^2}{2\beta}\mathbb{E}\Big(\big\|Y - \widetilde{m}\big\|^2\Big).\end{aligned}$$

Replacing in the previous inequalities, we obtain

Lemma 2.3 *With probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$,*

$$\begin{aligned}\langle\theta, \widehat{m} - \widetilde{m}\rangle = \frac{1}{n}\sum_{i=1}^n\langle\theta, Y_i - \widetilde{m}\rangle &\leq g_2(2\mu)\frac{\mu\lambda}{2}\mathbb{E}\big(\langle\theta, Y - \widetilde{m}\rangle^2\big) \\ &\quad + \exp(2\mu)g_1\left(\frac{2\mu^2}{\beta}\right)\frac{\mu\lambda}{2\beta}\mathbb{E}\big(\|Y - \widetilde{m}\|^2\big) + \frac{\beta + 2\log(\delta^{-1})}{2\mu\lambda n}.\end{aligned}$$

Let us put $a = g_2\big(2\mu\big)$ and $b = \exp(2\mu)g_1\left(\dfrac{2\mu^2}{\beta}\right)$ for short. Remark that

$$\begin{aligned}\langle\theta, Y - m\rangle^2 = \langle\theta, \alpha X - m\rangle^2 &= \Big(\alpha\langle\theta, X - m\rangle - (1-\alpha)\langle\theta, m\rangle\Big)^2 \\ &\leq \alpha\langle\theta, X - m\rangle^2 + (1-\alpha)\langle\theta, m\rangle^2 \leq \langle\theta, X - m\rangle^2 + (1-\alpha)\langle\theta, m\rangle^2.\end{aligned}$$

Therefore, using inequality (1) and the definition of α,

$$\mathbb{E}\big(\langle\theta, Y - \widetilde{m}\rangle^2\big) \leq \mathbb{E}\big(\langle\theta, Y - m\rangle^2\big) \leq \mathbb{E}\big(\langle\theta, X - m\rangle^2\big) + \langle\theta, m\rangle^2\inf_{p\geq 2}\frac{\lambda^p}{p+1}\left(\frac{p}{p+1}\right)^p\mathbb{E}\big(\|X\|^p\big).$$

In the same way,

$$\begin{aligned}\mathbb{E}\big(\|Y - \widetilde{m}\|^2\big) \leq \mathbb{E}\big(\|Y - m\|^2\big) &\leq \mathbb{E}\big(\alpha\|X - m\|^2 + (1-\alpha)\|m\|^2\big) \\ &\leq \mathbb{E}\big(\|X - m\|^2\big) + \|m\|^2\inf_{p\geq 2}\frac{\lambda^p}{p+1}\left(\frac{p}{p+1}\right)^p\mathbb{E}\big(\|X\|^p\big).\end{aligned}$$

In view of these remarks, the previous lemma translates to

Lemma 2.4 *With probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$,*

$$\begin{aligned}\langle\theta, \widehat{m} - m\rangle \leq \frac{a\mu\lambda}{2}\mathbb{E}\big(\langle\theta, X - m\rangle^2\big) + \frac{b\mu\lambda}{2\beta}\mathbb{E}\big(\|X - m\|^2\big) + \frac{\beta + 2\log(\delta^{-1})}{2\mu\lambda n} \\ + \inf_{p\geq 1}\frac{\lambda^p}{p+1}\left(\frac{p}{p+1}\right)^p\mathbb{E}\big(\|X\|^p\langle\theta, X - m\rangle_-\big) \\ + \inf_{p\geq 2}\frac{\lambda^p}{p+1}\left(\frac{p}{p+1}\right)^p\mathbb{E}\big(\|X\|^p\big)\left(\langle\theta, m\rangle_- + \frac{a\mu\lambda}{2}\langle\theta, m\rangle^2 + \frac{b\mu\lambda}{2\beta}\|m\|^2\right).\end{aligned}$$

Proposition 2.1 follows by optimizing the values of λ and β on the first three factors of the sum.

References

O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. pages 1–269.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré*, 48(4):1148–1185, 2012.

E. Joly, G. Lugosi, and R. I. Oliveira. On the estimation of the mean of a random vector. *Electronic Journal of Statistics*, 11:440–451, 2017.

G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics, to appear*, 2017.

S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 4:2308–2335, 2015.

S. Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Annals of Statistics, to appear*, 2016.

S. Minsker and X. Wei. Estimation of the covariance structure of heavy-tailed distributions. In *NIPS–2017, to appear*, 2017.