Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2012, Vol. 48, No. 4, 1148–1185
DOI: 10.1214/11-AIHP454
© Association des Publications de l'Institut Henri Poincaré, 2012

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

Challenging the empirical mean and empirical variance: A deviation study

Olivier Catoni[a]

[a]*CNRS – UMR 8553, Département de Mathématiques et Applications, Ecole Normale Supérieure, 45 rue d'Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team. E-mail: olivier.catoni@ens.fr*

Received 13 September 2010; revised 12 August 2011; accepted 18 August 2011

Abstract. We present new M-estimators of the mean and variance of real valued random variables, based on PAC-Bayes bounds. We analyze the non-asymptotic minimax properties of the deviations of those estimators for sample distributions having either a bounded variance or a bounded variance and a bounded kurtosis. Under those weak hypotheses, allowing for heavy-tailed distributions, we show that the worst case deviations of the empirical mean are suboptimal. We prove indeed that for any confidence level, there is some M-estimator whose deviations are of the same order as the deviations of the empirical mean of a Gaussian statistical sample, even when the statistical sample is instead heavy-tailed. Experiments reveal that these new estimators perform even better than predicted by our bounds, showing deviation quantile functions uniformly lower at all probability levels than the empirical mean for non-Gaussian sample distributions as simple as the mixture of two Gaussian measures.

Résumé. Nous présentons de nouveaux M-estimateurs de la moyenne et de la variance d'une variable aléatoire réelle, fondés sur des bornes PAC-Bayésiennes. Nous analysons les propriétés minimax non-asymptotiques des déviations de ces estimateurs pour des distributions de l'échantillon soit de variance bornée, soit de variance et de kurtosis bornées. Sous ces hypothèses faibles, permettant des distributions à queue lourde, nous montrons que les déviations de la moyenne empirique sont dans le pire des cas sous-optimales. Nous prouvons en effet que pour tout niveau de confiance, il existe un M-estimateur dont les déviations sont du même ordre que les déviations de la moyenne empirique d'un échantillon Gaussien, même dans le cas où la véritable distribution de l'échantillon a une queue lourde. Le comportement expérimental de ces nouveaux estimateurs est du reste encore meilleur que ce que les bornes théoriques laissent prévoir, montrant que la fonction quantile des déviations est constamment en dessous de celle de la moyenne empirique pour des échantillons non Gaussiens aussi simples que des mélanges de deux distributions Gaussiennes.

MSC: 62G05; 62G35

Keywords: Non-parametric estimation; M-estimators; PAC-Bayes bounds

1. Introduction

This paper is devoted to the estimation of the mean and possibly also of the variance of a real random variable from an independent identically distributed sample. While the most traditional way to deal with this question is to focus on the mean square error of estimators, we will instead focus on their deviations. Deviations are related to the estimation of confidence intervals which are of importance in many situations. While the empirical mean has an optimal minimax mean square error among all mean estimators in all models including Gaussian distributions, its deviations tell a different story. Indeed, as far as the mean square error is concerned, Gaussian distributions represent already the worst case, so that in the framework of a minimax mean least square analysis, no need is felt to improve estimators for non-Gaussian sample distributions. On the contrary, the deviations of estimators, and especially of the empirical mean, are worse for non-Gaussian samples than for Gaussian ones. Thus a deviation analysis will point

out possible improvements of the empirical mean estimator more successfully. It was nonetheless quite unexpected for us, and will undoubtedly also be for some of our readers, that the empirical mean could be improved, under such a weak hypothesis as the existence of a finite variance, and that this has remained unnoticed until now. One of the reasons may be that the weaknesses of the empirical mean disappear if we let the sample distribution be fixed and the sample size go to infinity. This does not mean however that a substantial improvement is not possible, nor that it is only concerned with specific sample sizes or weird worst case distributions: in the end of this paper, we will present experiments made on quite simple sample distributions, consisting in the mixture of two to three Gaussian measures, showing that more than twenty five percent can be gained on the widths of confidence intervals, for realistic sample sizes ranging from 100 to 2000. We think that, beyond the technicalities involved here, this exemplifies more broadly the pitfalls of asymptotic studies in statistics and should be quite thought provoking about the notions of optimality commonly used to assess the performance of estimators.

Our deviation study will use two kinds of tools: M-estimators to truncate observations and PAC-Bayesian theorems to combine estimates on the same sample without using a split scheme [1,2,8,12–14].

Its general conclusion is that, whereas the deviations of the empirical mean estimate may increase a lot when the sample distribution is far from being Gaussian, those of some new M-estimators will not. The improvement is the best for heavy-tailed distributions, as the worst case analysis performed to prove lower bounds will show. The improvement also increases as the confidence level at which deviations are computed increases.

Similar conclusions can be drawn in the case of least square regression with random design [3,4]. Discovering that using truncated estimators permits to get rid of sub Gaussian tail assumptions was the spur to study the simpler case of mean estimation for its own sake. Restricting the subject in this way (which is of course a huge restriction compared with least square regression) makes it possible to propose simpler dedicated estimators and to push their analysis further. It will indeed be possible here to obtain mathematical proofs for numerically significant non-asymptotic bounds.

The weakest hypothesis we will consider is the existence of a finite but unknown variance. In our M-estimators, adapting the truncation level depends on the value of the variance. However, this adaptation can be done without actually knowing the variance, through Lepski's approach.

Computing an observable confidence interval, on the other hand, requires more information. The simplest case is when the variance is known, or at least lower than a known bound. If it is not so, another possibility is to assume that the kurtosis is known, or lower than a known bound. Introducing the kurtosis is natural to our approach: in order to calibrate the truncation level for the estimation of the mean, we need to know the variance, and in the same way, in order to calibrate the truncation level for the estimation of the variance, we need to use the variance of the variance, which is provided by the kurtosis as a function of the variance itself.

In order to assess the quality of the results, we prove corresponding lower bounds for the best estimator confronted to the worst possible sample distribution, following the minimax approach. We also compute lower bounds for the deviations of the empirical mean estimator when the sample distribution is the worst possible. These latter bounds show the improvement that can be brought by M-estimators over the more traditional empirical mean. We plot the numerical values of these upper and lower bounds against the confidence level for typical finite sample sizes to show the gap between them.

The reader may wonder why we only consider the following extreme models, the narrow Gaussian model and the broad models

$$\mathcal{A}_{v_{\max}} = \left\{\mathbb{P} \in \mathcal{M}_+^1(\mathbb{R}):\ \mathbb{V}\mathrm{ar}_{\mathbb{P}} \leq v_{\max}\right\} \tag{1.1}$$

and

$$\mathcal{B}_{v_{\max},\kappa_{\max}} = \{\mathbb{P} \in \mathcal{A}_{v_{\max}}:\ \kappa_{\mathbb{P}} \leq \kappa_{\max}\}, \tag{1.2}$$

where $\mathbb{V}\mathrm{ar}_{\mathbb{P}}$ is the variance of $\mathbb{P}$, $\kappa_{\mathbb{P}}$ its kurtosis, and $\mathcal{M}_+^1(\mathbb{R})$ is the set of probability measures (positive measures of mass 1) on the real line equipped with the Borel sigma-algebra.

The reason is that, the minimax bounds obtained in these broad models being close to the ones obtained in the Gaussian model, introducing intermediate models would not change the order of magnitude of the bounds.

Let us end this introduction by advocating the value of confidence bounds, stressing more particularly the case of high confidence levels, since this is the situation where truncation brings the most decisive improvements.

One situation of interest which we will not comment further is when the estimated parameter is critical and making a big mistake on its value, even with a small probability, is unacceptable.

Another scenario to be encountered in statistical learning is the case when lots of estimates are to be computed and compared in the course of some decision making. Let us imagine, for instance, that some parameter $\theta \in \Theta$ is to be tuned in order to optimize the expected loss $\mathbb{E}[f_\theta(X)]$ of some family of loss functions $\{f_\theta\colon\ \theta \in \Theta\}$ computed on some random input X. Let us consider a split sample scheme where two i.i.d. samples $(X_1, \dots, X_s) \overset{\text{def}}{=} X_1^s$ and $(X_{s+1}, \dots, X_{s+n}) \overset{\text{def}}{=} X_{s+1}^{s+n}$ are used, one to build some estimators $\widehat{\theta}_k(X_1^s)$ of $\operatorname{argmin}_{\theta\in\Theta_k} \mathbb{E}[f_\theta(X)]$ in subsets Θ_k, $k = 1, \dots, K$, of Θ, and the other to test those estimators and keep hopefully the best. This is a very common model selection situation. One can think for instance of the choice of a basis to expand some regression function. If K is large, estimates of $\mathbb{E}[f_{\widehat{\theta}_k(X_1^s)}(X_{s+1})]$ will be required for a lot of values of k. In order to keep safe from over-fitting, very high confidence levels will be required if the resulting confidence level is to be computed through a union bound (because no special structure of the problem can be used to do better). Namely, a confidence level of $1-\varepsilon$ on the final result of the optimization on the test sample will require a confidence level of $1-\varepsilon/K$ for each mean estimate on the test sample. Even if ε is not very small (like, say, $5/100$), ε/K may be very small. For instance, if 10 parameters are to be selected among a set of 100, this gives $K = \binom{100}{10} \simeq 1.7 \cdot 10^{13}$. In practice, except in some special situations where fast algorithms exist, a heuristic scheme will be used to compute only a limited number of estimators $\widehat{\theta}_k$. An example of heuristics is to add greedily parameters one at a time, choosing at each step the one with the best estimated performance increase (in our example, this requires to compute 1000 estimators instead of $\binom{100}{10}$). Nonetheless, asserting the quality of the resulting choice requires a union bound on the whole set of possible outcomes of the data driven heuristic, and therefore calls for very high confidence levels for each estimate of the mean performance $\mathbb{E}[f_{\widehat{\theta}_k(X_1^s)}(X_{s+1})]$ on the test set.

The question we are studying in this paper should not be confused with robust statistics [9,10]. The most fundamental difference is that we are interested in estimating the mean of the sample distribution. In robust statistics, it is assumed that the sample distribution is in the neighbourhood of some known parametric model. This gives the possibility to replace the mean by some other location parameter, which as a rule will not be equal to the mean when the shape of the distribution is not constrained (and in particular is not assumed to be symmetric).

Other differences are that our point of view is non-asymptotic and that we study the deviations of estimators whereas robust statistics is focussed on their asymptotic mean square error.

Although we end up defining M-estimators with the help of influence functions, like in robust statistics, we use a truncation level depending on the sample size, whereas in robust statistics, the truncation level depends on the amount of contamination. Also, we truncate at much higher levels (that is we eliminate less outliers) that what would be advisable for instance in the case of a contaminated Gaussian statistical sample. Thus, although we have some tools in common with robust statistics, we use them differently to achieve a different purpose.

Adaptive estimation of a location parameter [5,6,15] is another setting where the empirical mean can be replaced by more efficient estimators. However, the setting studied by these authors is quite different from ours. The main difference, here again, is that the location parameter is assumed to be the center of symmetry of the sample distribution, a fact that is used to tailor location estimators based on a symmetrized density estimator. Another difference is that in these papers, the estimators are built with asymptotic properties in view, such as asymptotic normality with optimal variance and asymptotic robustness. These properties, although desirable, give no information on the non-asymptotic deviations we are studying here, and therefore do not provide as we do non-asymptotic confidence intervals.

2. Some new M-estimators

Let $(Y_i)_{i=1}^n$ be an i.i.d. sample drawn from some unknown probability distribution $\mathbb{P}$ on the real line $\mathbb{R}$ equipped with the Borel σ-algebra $\mathcal{B}$. Let Y be independent from $(Y_i)_{i=1}^n$ with the same marginal distribution $\mathbb{P}$. Assuming that $Y \in \mathbf{L}_2(\mathbb{P})$, let m be the mean of Y and let v be its variance:

$$\mathbb{E}(Y) = m \quad \text{and} \quad \mathbb{E}\big[(Y-m)^2\big] = v.$$

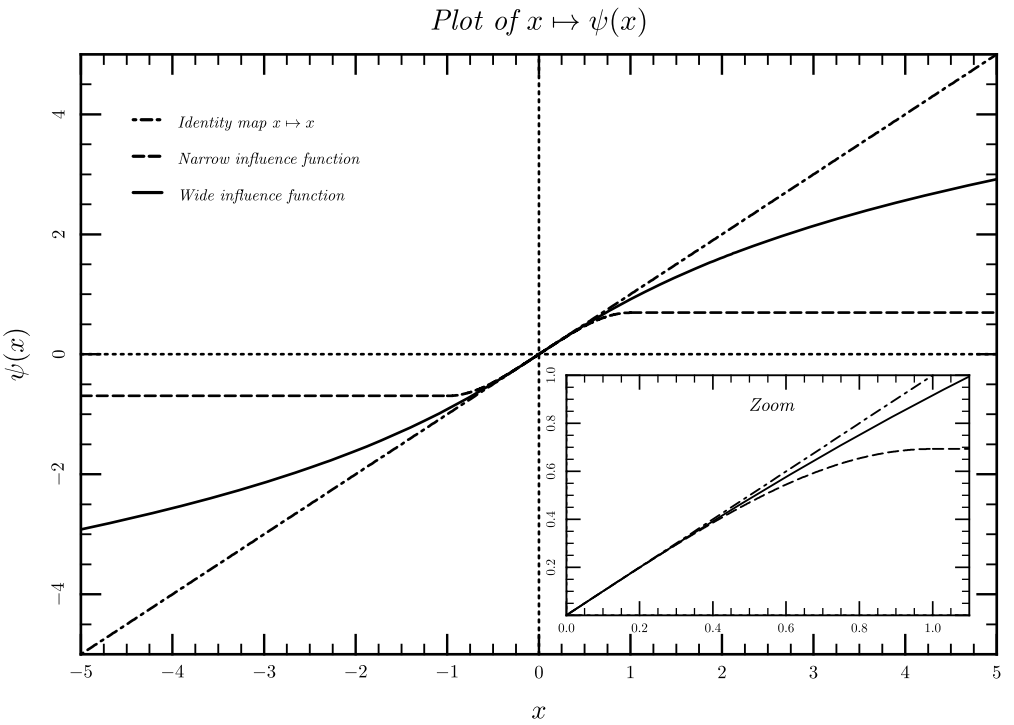

Fig. 1. Two possible choices of influence functions.

Let us consider some non-decreasing[1] influence function $\psi : \mathbb{R} \to \mathbb{R}$ such that

$$-\log\bigl(1 - x + x^2/2\bigr) \le \psi(x) \le \log\bigl(1 + x + x^2/2\bigr). \tag{2.1}$$

The *widest* possible choice of ψ (see Fig. 1) compatible with these inequalities is

$$\psi(x) = \begin{cases} \log\bigl(1 + x + x^2/2\bigr), & x \ge 0, \\ -\log\bigl(1 - x + x^2/2\bigr), & x \le 0, \end{cases} \tag{2.2}$$

whereas the *narrowest* possible choice is

$$\psi(x) = \begin{cases} \log(2), & x \ge 1, \\ -\log\bigl(1 - x + x^2/2\bigr), & 0 \le x \le 1, \\ \log\bigl(1 + x + x^2/2\bigr), & -1 \le x \le 0, \\ -\log(2), & x \le -1. \end{cases} \tag{2.3}$$

Although ψ is not the derivative of some explicit error function, we will use it in the same way, so that it can be considered as an influence function.

Indeed, α being some positive real parameter to be chosen later, we will build our estimator $\widehat{\theta}_\alpha$ of the mean m as the solution of the equation

$$\sum_{i=1}^{n} \psi\bigl[\alpha(Y_i - \widehat{\theta}_\alpha)\bigr] = 0.$$

[1] We would like to thank one of the anonymous referees of an early version of this paper for pointing out the benefits that could be drawn from the use of a non-decreasing influence function in this section.

(When the narrow choice of ψ defined by equation (2.3) is made, the above equation may have more than one solution, in which case any of them can be used to define $\widehat{\theta}_\alpha$.)

The widest choice of ψ is the one making $\widehat{\theta}_\alpha$ the closest to the empirical mean. For this reason it may be preferred if our aim is to stabilize the empirical mean by making the smallest possible change, which could be justified by the fact that the empirical mean is optimal in the case when the sample $(Y_i)_{i=1}^n$ is Gaussian.

Our analysis of $\widehat{\theta}_\alpha$ will rely on the following exponential moment inequalities, from which deviation bounds will follow. Let us introduce the quantity

$$r(\theta) = \frac{1}{\alpha n} \sum_{i=1}^n \psi\big[\alpha(Y_i - \theta)\big], \quad \theta \in \mathbb{R}.$$

Proposition 2.1.

$$\mathbb{E}\big\{\exp\big[\alpha n r(\theta)\big]\big\} \le \left\{1 + \alpha(m-\theta) + \frac{\alpha^2}{2}\big[v + (m-\theta)^2\big]\right\}^n$$
$$\le \exp\left\{n\alpha(m-\theta) + \frac{n\alpha^2}{2}\big[v + (m-\theta)^2\big]\right\}.$$

In the same way

$$\mathbb{E}\big\{\exp\big[-\alpha n r(\theta)\big]\big\} \le \left\{1 - \alpha(m-\theta) + \frac{\alpha^2}{2}\big[v + (m-\theta)^2\big]\right\}^n$$
$$\le \exp\left\{-n\alpha(m-\theta) + \frac{n\alpha^2}{2}\big[v + (m-\theta)^2\big]\right\}.$$

The proof of this proposition is an obvious consequence of inequalities (2.1) and of the fact that the sample is assumed to be i.i.d. It justifies the special choice of influence function we made. If we had taken for ψ the identity function, and thus for $\widehat{\theta}_\alpha$ the empirical mean, the exponential moments of $r(\theta)$ would have existed only in the case when the random variable Y itself has exponential moments. In order to bound $\widehat{\theta}_\alpha$, we will find two non-random values θ_- and θ_+ of the parameter such that with large probability $r(\theta_-) > 0 > r(\theta_+)$, which will imply that $\theta_- < \widehat{\theta}_\alpha < \theta_+$, since $r(\widehat{\theta}_\alpha) = 0$ by construction and $\theta \mapsto r(\theta)$ is non-increasing.

Proposition 2.2. *The values of the parameters $\alpha \in \mathbb{R}_+$ and $\varepsilon \in]0,1[$ being set, let us define for any $\theta \in \mathbb{R}$ the bounds*

$$B_+(\theta) = m - \theta + \frac{\alpha}{2}\big[v + (m-\theta)^2\big] + \frac{\log(\varepsilon^{-1})}{n\alpha},$$
$$B_-(\theta) = m - \theta - \frac{\alpha}{2}\big[v + (m-\theta)^2\big] - \frac{\log(\varepsilon^{-1})}{n\alpha}.$$

They satisfy

$$\mathbb{P}\big[r(\theta) < B_+(\theta)\big] \ge 1 - \varepsilon, \qquad \mathbb{P}\big[r(\theta) > B_-(\theta)\big] \ge 1 - \varepsilon.$$

The proof of this proposition is also straightforward: it is a mere consequence of Chebyshev's inequality and of the previous proposition. Let us assume that

$$\alpha^2 v + \frac{2\log(\varepsilon^{-1})}{n} \le 1.$$

Let θ_+ be the smallest solution of the quadratic equation $B_+(\theta_+) = 0$ and let θ_- be the largest solution of the equation $B_-(\theta_-) = 0$.

Lemma 2.3.

$$\theta_+ = m + \left(\frac{\alpha v}{2} + \frac{\log(\varepsilon^{-1})}{\alpha n}\right)\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - \alpha^2 v - \frac{2\log(\varepsilon^{-1})}{n}}\right)^{-1}$$
$$\leq m + \left(\frac{\alpha v}{2} + \frac{\log(\varepsilon^{-1})}{\alpha n}\right)\left(1 - \frac{\alpha^2 v}{2} - \frac{\log(\varepsilon^{-1})}{n}\right)^{-1},$$
$$\theta_- = m - \left(\frac{\alpha v}{2} + \frac{\log(\varepsilon^{-1})}{\alpha n}\right)\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - \alpha^2 v - \frac{2\log(\varepsilon^{-1})}{n}}\right)^{-1}$$
$$\geq m - \left(\frac{\alpha v}{2} + \frac{\log(\varepsilon^{-1})}{\alpha n}\right)\left(1 - \frac{\alpha^2 v}{2} - \frac{\log(\varepsilon^{-1})}{n}\right)^{-1}.$$

Moreover, since the map $\theta \mapsto r(\theta)$ is non-increasing, with probability at least $1 - 2\varepsilon$,

$$\theta_- < \widehat{\theta}_\alpha < \theta_+.$$

The proof of this lemma is also an obvious consequence of the previous proposition and of the definitions of θ_+ and θ_-. Optimizing the choice of α provides

Proposition 2.4. *Let us assume that $n > 2\log(\varepsilon^{-1})$ and let us consider*

$$\eta = \sqrt{\frac{2v\log(\varepsilon^{-1})}{n(1 - 2\log(\varepsilon^{-1})/n)}} \quad \text{and} \quad \alpha = \sqrt{\frac{2\log(\varepsilon^{-1})}{n(v + \eta^2)}}.$$

In this case $\theta_+ = m + \eta$ and $\theta_- = m - \eta$, so that with probability at least $1 - 2\varepsilon$,

$$|m - \widehat{\theta}_\alpha| \leq \eta = \sqrt{\frac{2v\log(\varepsilon^{-1})}{n(1 - 2\log(\varepsilon^{-1})/n)}}.$$

In the same way, if we want to make a choice of α independent from ε, we can choose

$$\alpha = \sqrt{\frac{2}{nv}},$$

and assume that

$$n > 2\big[1 + \log\big(\varepsilon^{-1}\big)\big].$$

In this latter case, with probability at least $1 - 2\varepsilon$,

$$|\widehat{\theta}_\alpha - m| \leq \frac{1 + \log(\varepsilon^{-1})}{1/2 + (1/2)\sqrt{1 - 2[1 + \log(\varepsilon^{-1})]/n}}\sqrt{\frac{v}{2n}} \leq \frac{1 + \log(\varepsilon^{-1})}{1 - (1 + \log(\varepsilon^{-1}))/n}\sqrt{\frac{v}{2n}}.$$

In Figs 2–4, we compare the bounds on the deviations of our M-estimator $\widehat{\theta}_\alpha$ with the deviations of the empirical mean $M = \frac{1}{n}\sum_{i=1}^n Y_i$ when the sample distribution is Gaussian and when it belongs to the model $\mathcal{A}_1$ defined in the introduction by equation (1.1), page 1149. (The bounds for the empirical mean will be explained and proved in subsequent sections.)

More precisely, the deviation upper bounds for our estimator $\widehat{\theta}_\alpha$ for the worst sample distribution in $\mathcal{A}_1$, the model defined by (1.1), page 1149, are compared with the exact deviations of the empirical mean M of a Gaussian sample. This is the minimax bound at all confidence levels in the Gaussian model, as will be proved later on. Consequently, the deviations of our estimator cannot be smaller for the worst sample distribution in $\mathcal{A}_1$, which contains Gaussian

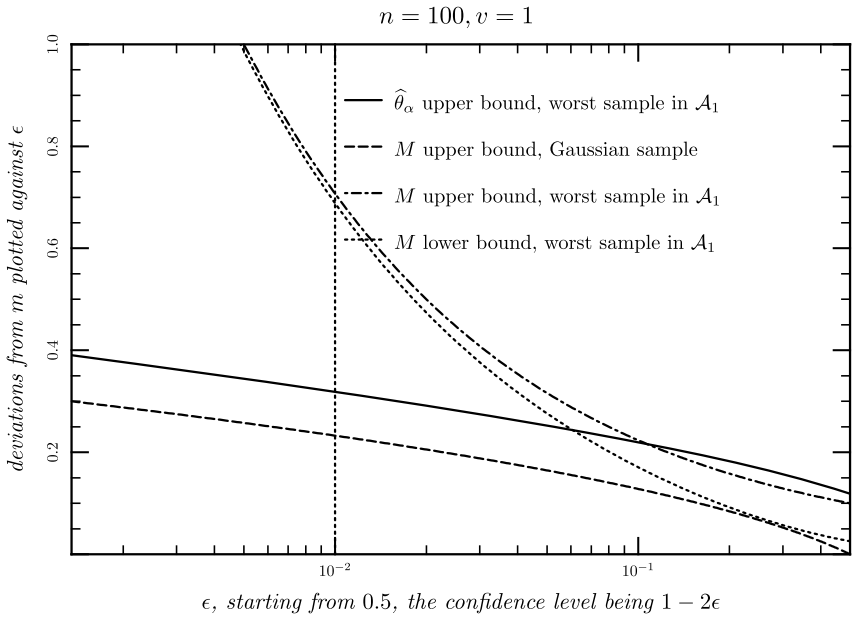

Fig. 2. Deviations of $\widehat{\theta}_\alpha$ from the sample mean, compared with those of the empirical mean.

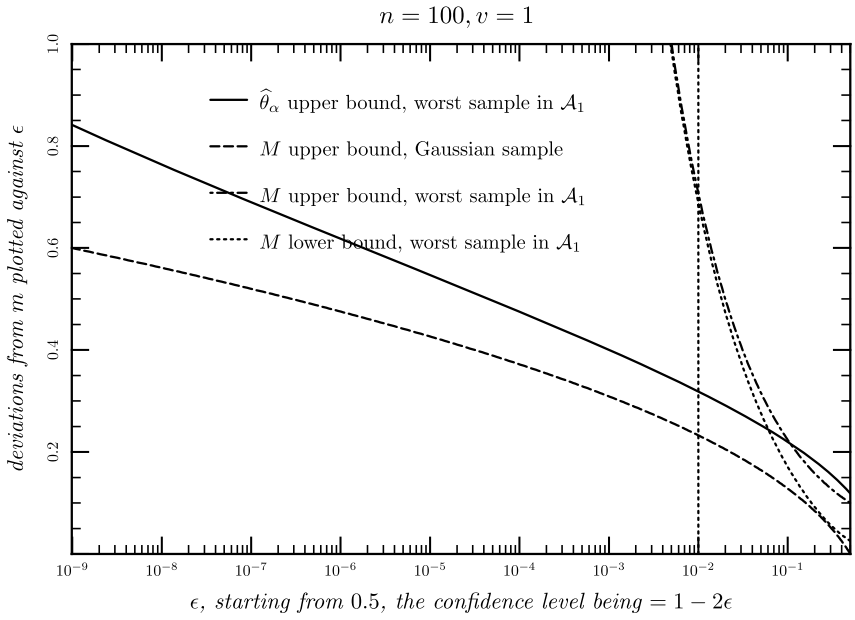

Fig. 3. Same as Fig. 2, but showing a larger range of confidence levels.

distributions. We see on the plots that starting from $\varepsilon = 0.1$ (corresponding to a 80% confidence level), our upper bound is close to being minimax, not only in $\mathcal{A}_1$, but also in the small Gaussian sub-model. This shows that the deviations of our estimator are close to reach the minimax bound in any intermediate model containing Gaussian distributions and contained in $\mathcal{A}_1$.

Our estimator is also compared with the deviations of the empirical mean for the worst distribution in $\mathcal{A}_1$ (to be established later). In particular the lower bound proves that there are sample distributions in $\mathcal{A}_1$ for which the

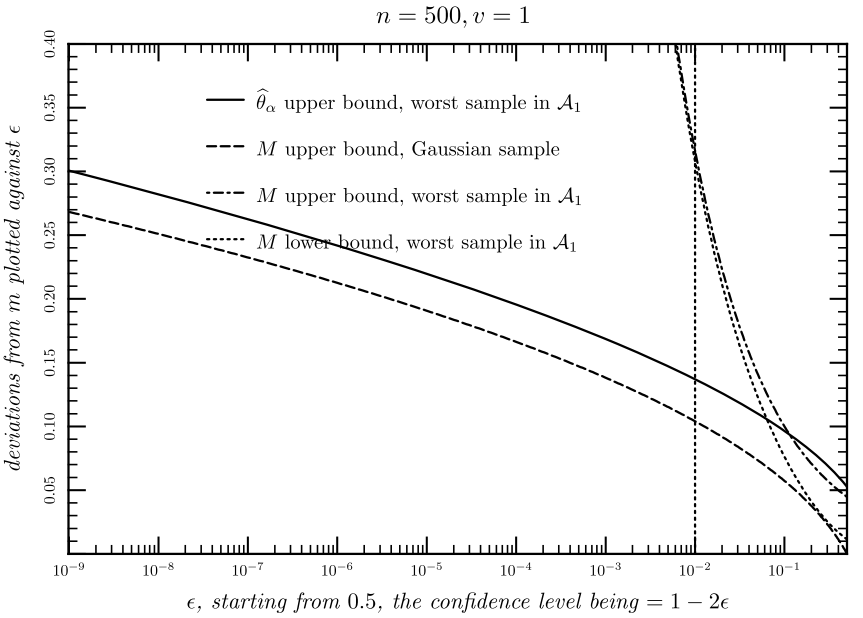

Fig. 4. Deviations of $\widehat{\theta}_\alpha$ for a larger sample size.

deviations of the empirical mean are far from being optimal, showing the need to introduce a new estimator to correct this.

In Fig. 2, we chose a sample size $n = 100$ and plotted the deviations against the confidence level (or rather against ε, the confidence level being $1 - 2\varepsilon$). As shown in Fig. 3, showing a wider range of ε values, our bound stays close to the Gaussian bound up to very high confidence levels (up to $\varepsilon = 10^{-9}$ and more). On the other hand, it already outperforms the empirical mean by a factor larger than two at confidence level 98% (that is for $\varepsilon = 10^{-2}$).

When we increase the sample size to $n = 500$, as in Fig. 4, the performance of our M-estimator is even closer to optimal.

3. Adapting to an unknown variance

In this section, we will use Lepski's renowned adaptation method [11] when nothing is known, except that the variance is finite. Under so uncertain (but unfortunately so frequent) circumstances, it is impossible to provide any *observable* confidence intervals, but it is still possible to define an adaptive estimator and to bound its deviations by unobservable bounds (depending on the unknown variance). To understand the subject of this section, one should keep in mind that *adapting* to the variance is a weaker requirement than *estimating* the variance: estimating the variance at any predictable rate would require more prior information (bearing for instance on higher moments of the sample distribution).

The idea of Lepski's method is powerful and simple: consider a sequence of confidence intervals obtained by assuming that the variance is bounded by a sequence of bounds v_k and pick up as an estimator the middle of the smallest interval intersecting all the larger ones. For this to be legitimate, we need all the confidence regions for which the variance bound is valid to hold together, which is performed using a union bound.

Let us describe this idea more precisely. Let $\widehat{\theta}(v_{\max})$ be some estimator of the mean depending on some assumed variance bound $v_{\max}$, as the ones described in the beginning of this paper. Let $\delta(v_{\max}, \varepsilon) \in \mathbb{R}_+ \cup \{+\infty\}$ be some deviation bound[2] proved in $\mathcal{A}_{v_{\max}}$: namely let us assume that for any sample distribution in $\mathcal{A}_{v_{\max}}$, with probability at

[2] $\mathcal{A}_{v_{\max}}$ is defined by (1.1), page 1149.

least $1-2\varepsilon$,

$$\bigl|m-\widehat{\theta}(v_{\max})\bigr| \le \delta(v_{\max},\varepsilon).$$

Let us also decide by convention that $\delta(v_{\max},0)=+\infty$.

Let $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$ be some coding atomic probability measure on the positive real line, which will serve to take a union bound on a (countable) set of possible values of $v_{\max}$.

We can choose for instance for ν the following coding distribution: expressing $v_{\max}$ by comparison with some reference value V,

$$v_{\max} = V2^s \sum_{k=0}^{d} c_k 2^{-k}, \quad s\in\mathbb{Z}, d\in\mathbb{N}, (c_k)_{k=0}^d \in \{0,1\}^{d+1}, c_0=c_d=1,$$

we set

$$\nu(v_{\max}) = \frac{2^{-2(d-1)}}{5(|s|+2)(|s|+3)}$$

and otherwise we set $\nu(v_{\max})=0$. It is easy to see that this defines a probability distribution on $\mathbb{R}_+$ (supported by dyadic numbers scaled by the factor V). It is clear that, as far as possible, the reference value V should be chosen as close as possible to the true variance v.

Another possibility is to set for some parameters $V\in\mathbb{R}$, $\rho>1$ and $s\in\mathbb{N}$,

$$\nu\bigl(V\rho^{2k}\bigr) = \frac{1}{2s+1}, \quad k\in\mathbb{Z}, |k|\le s. \tag{3.1}$$

Let us consider for any $v_{\max}$ such that $\delta(v_{\max},\varepsilon\nu(v_{\max}))<+\infty$ the confidence interval

$$I(v_{\max}) = \widehat{\theta}(v_{\max}) + \delta\bigl[v_{\max},\varepsilon\nu(v_{\max})\bigr]\times(-1,1).$$

Let us put $I(v_{\max})=\mathbb{R}$ when $\delta(v_{\max},\varepsilon\nu(v_{\max}))=+\infty$.

Let us consider the non-decreasing family of closed intervals

$$J(v_1) = \bigcap\bigl\{I(v_{\max}):\ v_{\max}\ge v_1\bigr\}, \quad v_1\in\mathbb{R}_+.$$

(In this definition, we can restrict the intersection to the support of ν, since otherwise $I(v_{\max})=\mathbb{R}$.) A union bound shows immediately that with probability at least $1-2\varepsilon$, $m\in J(v)$, implying as a consequence that $J(v)\ne\varnothing$.

Proposition 3.1. *Since $v_1\mapsto J(v_1)$ is a non-decreasing family of closed intervals, the intersection*

$$\bigcap\bigl\{J(v_1): v_1\in\mathbb{R}_+, J(v_1)\ne\varnothing\bigr\}$$

is a non-empty closed interval, and we can therefore pick up an adaptive estimator $\widetilde{\theta}$ belonging to it, choosing for instance the middle of this interval.

With probability at least $1-2\varepsilon$, $m\in J(v)$, which implies that $J(v)\ne\varnothing$, and therefore that $\widetilde{\theta}\in J(v)$.

Thus with probability at least $1-2\varepsilon$

$$|m-\widetilde{\theta}| \le \bigl|J(v)\bigr| \le 2\inf_{v_{\max}>v} \delta\bigl(v_{\max},\varepsilon\nu(v_{\max})\bigr).$$

If the confidence bound $\delta(v_{\max},\varepsilon)$ is homogeneous, in the sense that

$$\delta(v_{\max},\varepsilon) = B(\varepsilon)\sqrt{v_{\max}},$$

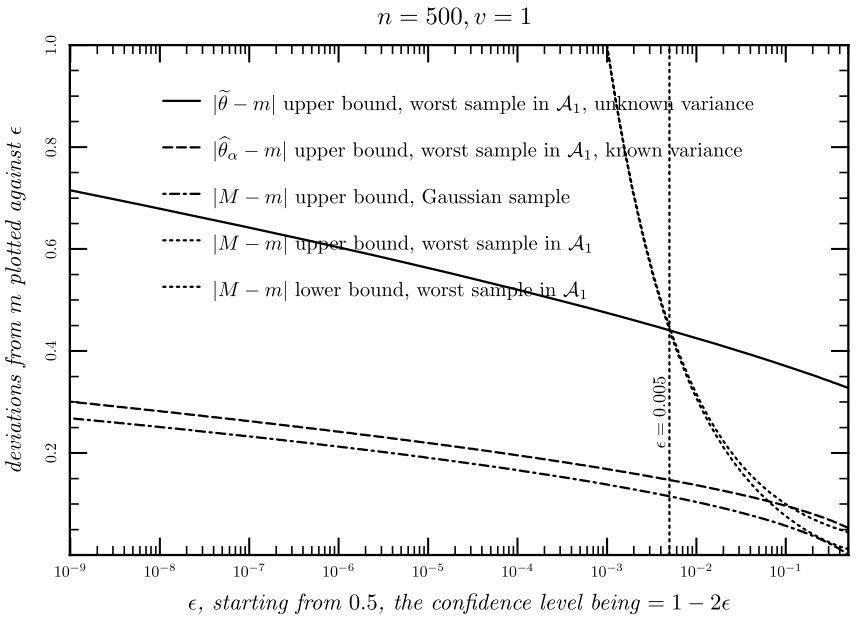

Fig. 5. Deviations of $\widetilde{\theta}$, the adaptive mean estimator for a sample with unknown variance, compared with other estimators.

as it is the case in Proposition 2.4 (*page* 1153) *with*

$$B(\varepsilon) = \sqrt{\frac{2\log(\varepsilon^{-1})}{n(1 - 2\log(\varepsilon^{-1})/n)}}$$

then with probability at least $1 - 2\varepsilon$,

$$|m - \widetilde{\theta}| \leq 2 \inf_{v_{\max} > v} B\big[\varepsilon\, \nu(v_{\max})\big]\sqrt{v_{\max}}.$$

Thus in the case when ν *is defined by equation* (3.1), *page* 1156, *and* $|\log(v/V)| \leq 2s\log(\rho)$, *with probability at least* $1 - 2\varepsilon$

$$|m - \widetilde{\theta}| \leq 2\rho B\left(\frac{\varepsilon}{2s+1}\right)\sqrt{v}.$$

Let us see what happens for a sample size of $n = 500$, when we assume that $|\log(v/V)| \leq 2\log(100)$ and we take $\rho = 1.05$.

Figure 5 shows that, for a sample of size $n = 500$, there are sample distributions with a finite variance for which the deviations of the empirical mean blow up for confidence levels higher than 99%, where as the deviations of our adaptive estimator remain under control, even at confidence levels as high as $1 - 10^{-9}$.

The conclusion is that if our aim is to minimize the worst estimation error over 100 statistical experiments or more and we have no information on the standard deviation except that it is in some range of the kind $(1/100, 100)$ (which is pretty huge and could be increased even more if desired), then the performance of the empirical mean estimator for the worst sample distribution breaks down but thresholding very large outliers as $\widetilde{\theta}$ does can cure this problem.

4. Mean and variance estimates depending on the kurtosis

Situations where the variance is unknown are likely to happen. We have seen in the previous section how to adapt to an unknown variance. The price to pay is a loss of a factor two in the deviation bound, and the fact that it is no longer observable.

Here we will make hypotheses under which it is possible to estimate both the variance and the mean, and to obtain an observable confidence interval, without loosing a factor two as in the previous section. Making use of the kurtosis parameter is the most natural way to achieve these goals in the framework of our approach. This is what we are going to do here.

4.1. *Some variance estimate depending on the kurtosis*

In this section, we are going to consider an alternative to the unbiased usual variance estimate

$$\begin{aligned}\widehat{V} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j\right)^2 \\ &= \frac{1}{n(n-1)}\sum_{1\le i<j\le n}(Y_i - Y_j)^2. \end{aligned} \tag{4.1}$$

We will assume that the fourth moment $\mathbb{E}(Y^4)$ is finite and that some upper bound is known for the kurtosis

$$\kappa = \frac{\mathbb{E}[(Y-m)^4]}{v^2}.$$

Our aim will be, as before with the mean, to define an estimate with better deviations than $\widehat{V}$. We will use κ in the following computations, but when only an upper bound is known, κ can be replaced with this upper bound in the definition of the estimator and the estimates of its performance.

Let us write $n = pq + r$, with $0 \le r < p$, and let $\{1,\dots,n\} = \bigsqcup_{\ell=1}^{q} I_\ell$ be the partition of the n first integers defined by

$$I_\ell = \begin{cases} \{i\in\mathbb{N};\ p(\ell-1) < i \le p\ell\}, & 1\le \ell < q, \\ \{i\in\mathbb{N};\ p(\ell-1) < i \le n\}, & \ell = q. \end{cases}$$

We will develop some kind of block threshold scheme, introducing

$$\begin{aligned} Q_\delta(\beta) &= \frac{1}{q}\sum_{\ell=1}^{q}\psi\left(\frac{1}{|I_\ell|(|I_\ell|-1)}\sum_{\substack{i,j\in I_\ell\\ i<j}}\left[\beta(Y_i-Y_j)^2 - 2\delta\right]\right) \\ &= \frac{1}{q}\sum_{\ell=1}^{q}\psi\left[\frac{\beta}{|I_\ell|-1}\sum_{i\in I_\ell}\left(Y_i - \frac{1}{|I_\ell|}\sum_{j\in I_\ell} Y_j\right)^2 - \delta\right], \end{aligned}$$

where ψ is a non-decreasing influence function satisfying (2.1), page 1151.

If ψ were replaced with the identity, we would have $\mathbb{E}[Q_\delta(\beta)] = \beta v - \delta$. The idea is to solve $Q_\delta(\widehat{\beta}) = 0$ in $\widehat{\beta}$ and to estimate v by $\delta/\widehat{\beta}$. Anyhow, for technical reasons, we will adopt a slightly different definition for $\widehat{\beta}$ as well as for the estimate of v, as we will show now. Let us first get some deviation inequalities for $Q_\delta(\beta)$, derived as usual from exponential bounds. It is straightforward to see that

$$\begin{aligned}\mathbb{E}\left\{\exp\left[q\,Q(\beta)\right]\right\} \le \prod_{\ell=1}^{q}\Bigg\{1 + (\beta v - \delta) \\ + \frac{1}{2}\left[(\beta v-\delta)^2 + \frac{\beta^2}{|I_\ell|^2(|I_\ell|-1)^2}\sum_{\substack{i<j\in I_\ell\\ s<t\in I_\ell}}\mathbb{E}\left\{\left[(Y_i-Y_j)^2 - 2v\right]\left[(Y_s-Y_t)^2 - 2v\right]\right\}\right]\Bigg\}.\end{aligned}$$

We can now compute for any $i \neq j$

$$\begin{aligned}\mathbb{E}\big\{\big[(Y_i - Y_j)^2 - 2v\big]^2\big\} &= \mathbb{E}\big[(Y_i - Y_j)^4\big] - 4v^2 \\ &= 2\kappa v^2 + 6v^2 - 4v^2 = 2(\kappa+1)v^2,\end{aligned}$$

and for any distinct values of i, j and s,

$$\begin{aligned}&\mathbb{E}\big\{\big[(Y_i - Y_s)^2 - 2v\big]\big[(Y_j - Y_s)^2 - 2v\big]\big\} \\ &\quad= \mathbb{E}\big[(Y_i - Y_s)^2(Y_j - Y_s)^2\big] - 4v^2 \\ &\quad= \mathbb{E}\big\{\big[(Y_i - m)^2 + 2(Y_i - m)(Y_s - m) + (Y_s - m)^2\big] \\ &\qquad\times \big[(Y_j - m)^2 + 2(Y_j - m)(Y_s - m) + (Y_s - m)^2\big]\big\} - 4v^2 \\ &\quad= \mathbb{E}\big[(Y_s - m)^4\big] + \mathbb{E}\big\{(Y_s - m)^2\big[(Y_i - m)^2 + (Y_j - m)^2\big]\big\} + \mathbb{E}\big[(Y_i - m)^2(Y_j - m)^2\big] - 4v^2 \\ &\quad= (\kappa - 1)v^2.\end{aligned}$$

Thus

$$\begin{aligned}&\sum_{\substack{i<j\in I_\ell \\ s<t\in I_\ell}} \mathbb{E}\big\{\big[(Y_i - Y_j)^2 - 2v\big]\big[(Y_s - Y_t)^2 - 2v\big]\big\} \\ &\quad= |I_\ell|\big(|I_\ell| - 1\big)(\kappa+1)v^2 + |I_\ell|\big(|I_\ell| - 1\big)\big(|I_\ell| - 2\big)(\kappa - 1)v^2 \\ &\quad= |I_\ell|\big(|I_\ell| - 1\big)^2\left[(\kappa - 1) + \frac{2}{|I_\ell| - 1}\right]v^2.\end{aligned}$$

It shows that

$$\mathbb{E}\big\{\exp\big[q\,Q(\beta)\big]\big\} \leq \left\{1 + (\beta v - \delta) + \frac{1}{2}(\beta v - \delta)^2 + \frac{\beta^2 v^2}{2p}\left[\kappa - 1 + \frac{2}{p-1}\right]\right\}^q.$$

In the same way

$$\mathbb{E}\big\{\exp\big[-q\,Q(\beta)\big]\big\} \leq \left\{1 - (\beta v - \delta) + \frac{1}{2}(\beta v - \delta)^2 + \frac{\beta^2 v^2}{2p}\left[\kappa - 1 + \frac{2}{p-1}\right]\right\}^q.$$

Let $\chi = \kappa - 1 + \frac{2}{p-1} \overset{p\to\infty}{\simeq} \kappa - 1$. For any given $\beta_1, \beta_2 \in \mathbb{R}_+$, with probability at least $1 - 2\varepsilon_1$,

$$Q(\beta_1) < \beta_1 v - \delta + \frac{1}{2}(\beta_1 v - \delta)^2 + \frac{\chi\beta_1^2 v^2}{2p} + \frac{\log(\varepsilon_1^{-1})}{q},$$

$$Q(\beta_2) > \beta_2 v - \delta - \frac{1}{2}(\beta_2 v - \delta)^2 - \frac{\chi\beta_2^2 v^2}{2p} - \frac{\log(\varepsilon_1^{-1})}{q}.$$

Let us define, for some parameter $y \in \mathbb{R}$, $\widehat{\beta}$ as

$$Q(\widehat{\beta}) = -y,$$

and let us choose

$$y = \frac{\chi\delta^2}{2p} + \frac{\log(\varepsilon_1^{-1})}{q} \quad \text{and} \quad \beta_2 = \frac{\delta}{v},$$

so that $Q(\beta_2) > -y$. Let us put $\xi = \delta - \beta_1 v$ and let us choose ξ such that $Q(\beta_1) < -y$. This implies that ξ is solution of

$$\frac{1+\zeta}{2}\xi^2 - (1+\zeta\delta)\xi + 2y \le 0, \quad \text{where } \zeta = \frac{\chi}{p}.$$

Provided that $(1+\zeta\delta)^2 \ge 4(1+\zeta)y$, the smallest solution of this equation is

$$\xi = \frac{4y}{1+\zeta\delta + \sqrt{(1+\zeta\delta)^2 - 4(1+\zeta)y}}.$$

With these parameters, with probability at least $1-2\varepsilon_1$, $Q[(\delta-\xi)/v] < -y < Q(\delta/v)$, implying that

$$\frac{\delta-\xi}{v} \le \widehat{\beta} \le \frac{\delta}{v}.$$

Thus putting

$$\widehat{v} = \frac{\sqrt{\delta(\delta-\xi)}}{\widehat{\beta}},$$

we get

$$\left(1-\frac{\xi}{\delta}\right)^{1/2} \le \frac{v}{\widehat{v}} \le \left(1-\frac{\xi}{\delta}\right)^{-1/2}.$$

In order to minimize $\frac{\xi}{\delta} \simeq \frac{2y}{\delta}$, we are led to take $\delta = \sqrt{\frac{2p\log(\varepsilon_1^{-1})}{\chi q}}$. We get

$$y = \frac{2\log(\varepsilon_1^{-1})}{q}, \qquad \frac{y}{\delta} = \sqrt{\frac{2\chi\log(\varepsilon_1^{-1})}{n-r}},$$

$$\zeta\delta = \frac{y}{\delta} \quad \text{and} \quad (1+\zeta)y = \frac{2\log(\varepsilon_1^{-1})}{q} + \frac{2\chi\log(\varepsilon_1^{-1})}{n-r}.$$

Thus the condition becomes

$$q \ge 8\log\bigl(\varepsilon_1^{-1}\bigr)\left(1+\frac{\chi}{p}\right)\left(1+\sqrt{\frac{2\chi\log(\varepsilon_1^{-1})}{n-r}}\right)^{-2}. \tag{4.2}$$

Proposition 4.1. *Under condition* (4.2), *with probability at least* $1-2\varepsilon_1$,

$$\bigl|\log(v) - \log(\widehat{v})\bigr| \le -\frac{1}{2}\log\left(1-\frac{\xi}{\delta}\right).$$

A simpler result is obtained by choosing $\xi = 2y(1+2y)$ (the values of y and δ being kept the same, so that we modify only the choice of β_1 through a different choice of ξ). In this case, $Q(\beta_1) < -y$ as soon as

$$(1+2y)^2 \le \frac{2+2\zeta\delta+\zeta\delta/y}{1+\zeta} = 2 + \frac{2\zeta\delta+\zeta\delta/y-2\zeta}{1+\zeta},$$

which is true as soon as $(1+2y)^2 \le 2$ and $\frac{\delta}{y} \ge 2$, yielding the simplified condition

$$\log\bigl(\varepsilon_1^{-1}\bigr) \le \min\left\{\frac{q}{4(1+\sqrt{2})}, \frac{n-r}{8\chi}\right\}. \tag{4.3}$$

In this case, we get with probability at least $1-2\varepsilon_1$ that

$$\left|\log(v)-\log(\widehat{v})\right| \le -\frac{1}{2}\log\left(1-\frac{2y(1+2y)}{\delta}\right) \simeq \frac{y}{\delta}.$$

Proposition 4.2. *Under condition* (4.3), *with probability at least* $1-2\varepsilon_1$,

$$\begin{aligned}\left|\log(v)-\log(\widehat{v})\right| &\le -\frac{1}{2}\log\left[1-2\sqrt{\frac{2\chi\log(\varepsilon_1^{-1})}{n-r}}\left(1+\frac{4\log(\varepsilon_1^{-1})}{q}\right)\right]\\ &\simeq \sqrt{\frac{2\chi\log(\varepsilon_1^{-1})}{n}}.\end{aligned}$$

Recalling that $\chi=\kappa-1+\frac{2}{p-1}$, we can choose in the previous proposition the approximately optimal block size

$$p=\left\lfloor\sqrt{\frac{n}{(\kappa-1)[4\log(\varepsilon_1^{-1})+1/2]}}\right\rfloor.$$

Corollary 4.3. *For this choice of parameter, as soon as the kurtosis* (*or its upper bound*) $\kappa\ge 3$, *under the condition*

$$\log\left(\varepsilon_1^{-1}\right) \le \frac{n}{36(\kappa-1)}-\frac{1}{8} \tag{4.4}$$

with probability at least $1-2\varepsilon_1^{-1}$,

$$\begin{aligned}&\left|\log(v)-\log(\widehat{v})\right|\\ &\quad\le -\frac{1}{2}\log\left[1-2\sqrt{\frac{2(\kappa-1)\log(\varepsilon_1^{-1})}{n}}\exp\left(4\sqrt{\frac{4\log(\varepsilon_1^{-1})+1/2}{(\kappa-1)n}}\right)\right]\\ &\quad\overset{n\to\infty}{\simeq} \sqrt{\frac{2(\kappa-1)\log(\varepsilon_1^{-1})}{n}}.\end{aligned}$$

This is the asymptotics we hoped for, since the variance of $(Y_i-m)^2$ is equal to $(\kappa-1)v^2$. The proof starts on page 1172.

Let us plot some curves, showing the tighter bounds of Proposition 4.1 (page 1160), with optimal choice of p (Figs 6–8). We compare our deviation bounds with the exact deviation quantiles of the variance estimate of equation (4.1), page 1158, applied to a Gaussian sample (given by a χ^2 distribution). This demonstrates that we can stay of the same order under much weaker assumptions.

4.2. *Mean estimate under a kurtosis assumption*

Here, we are going to plug a variance estimate $\widehat{v}$ into a mean estimate. Let us therefore assume that $\widehat{v}$ is a variance estimate such that with probability at least $1-2\varepsilon_1$,

$$\left|\log(v)-\log(\widehat{v})\right| \le \zeta.$$

This estimate may for example be the one defined in the previous section. Let $\widehat{\alpha}$ be some estimate of the desired value of the parameter α, to be defined later as a function of $\widehat{v}$. Let us define $\widehat{\theta}=\widehat{\theta}_{\widehat{\alpha}}$ by

$$r(\widehat{\theta}) = \frac{1}{n\widehat{\alpha}}\sum_{i=1}^{n}\psi\left[\widehat{\alpha}(Y_i-\widehat{\theta})\right]=0, \tag{4.5}$$

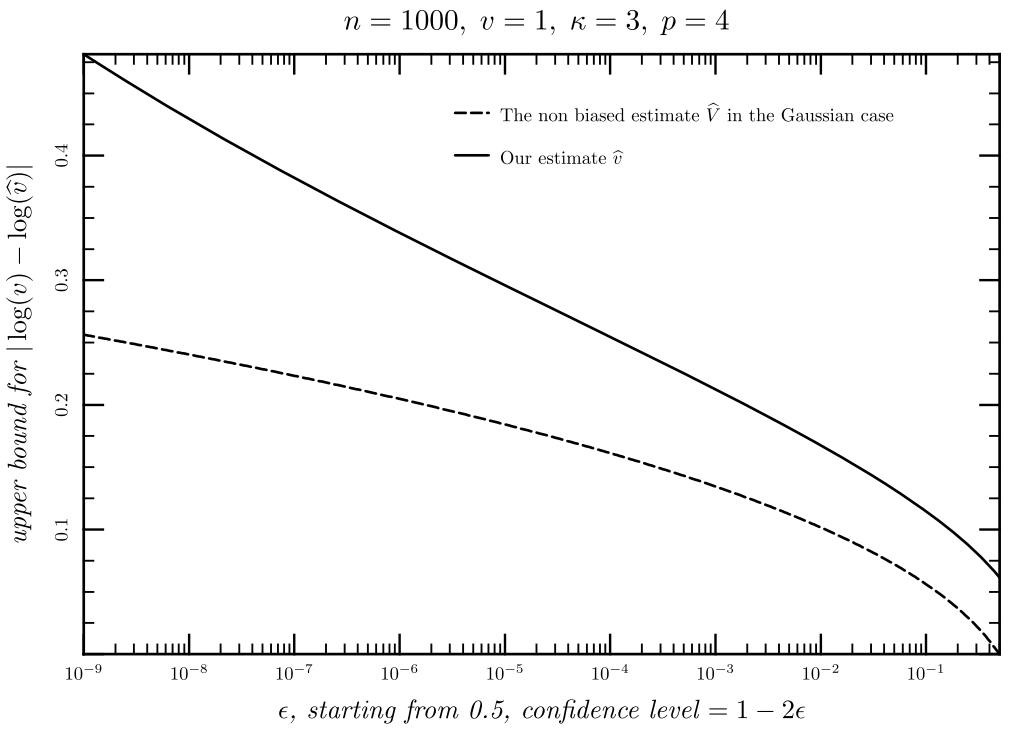

Fig. 6. Comparison of variance estimates.

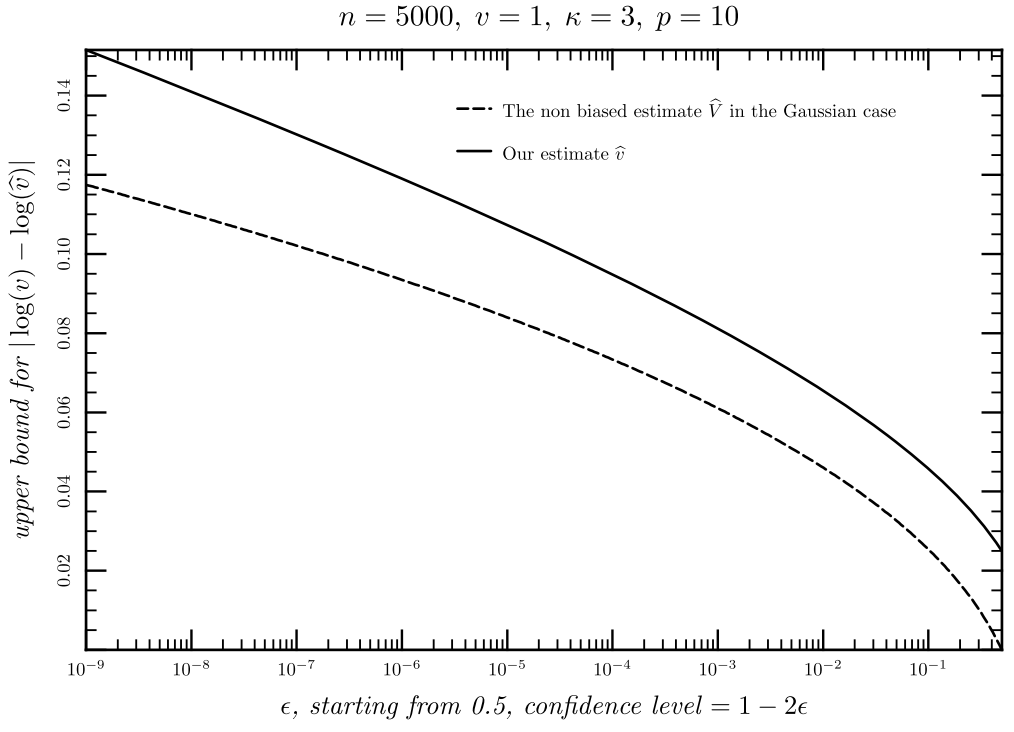

Fig. 7. Variance estimates for a larger sample size.

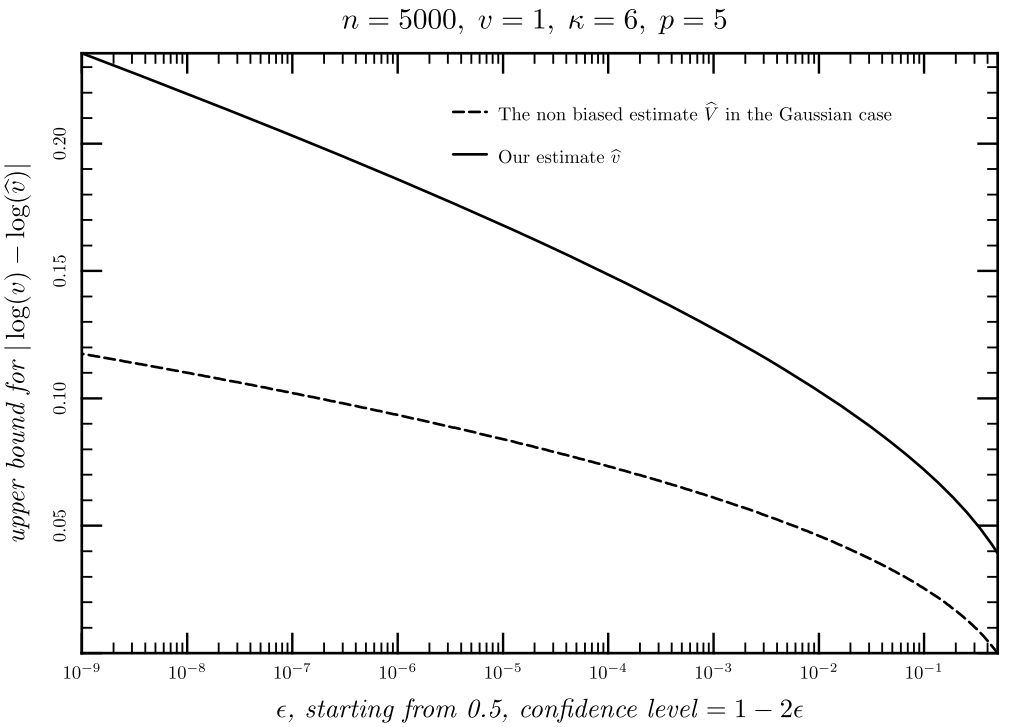

Fig. 8. Variance estimates for a larger kurtosis.

where ψ is the *narrow* influence function defined by equation (2.3), page 1151. As usual, we are looking for non-random values θ_- and θ_+ such that with large probability $r(\theta_+) < 0 < r(\theta_-)$, implying that $\theta_- < \widehat{\theta} < \theta_+$. But there is a new difficulty, caused by the fact that $\widehat{\alpha}$ will be an estimate depending on the value of the sample. This problem will be solved with the help of PAC-Bayes inequalities.

To take advantage of PAC-Bayes theorems, we are going to compare $\widehat{\alpha}$ with a perturbation $\widetilde{\alpha}$ built with the help of some supplementary random variable. Let indeed U be some uniform real random variable on the interval $(-1, +1)$, independent from everything else. Let us consider

$$\widetilde{\alpha} = \widehat{\alpha} + x\alpha \sinh(\zeta/2)U.$$

Let ρ be the distribution of $\widetilde{\alpha}$ given the sample value. We are going to compare this *posterior* distribution (meaning that it depends on the sample), with a *prior* distribution that is independent from the sample. Let π be the uniform probability distribution on the interval $(\alpha[\exp(-\zeta/2) - x\sinh(\zeta/2)], \alpha[\exp(\zeta/2) + x\sinh(\zeta/2)])$. Let us assume that $\widehat{\alpha}$ and α are defined with the help of some positive constant c as $\widehat{\alpha} = \sqrt{\frac{c}{\widehat{v}}}$ and $\alpha = \sqrt{\frac{c}{v}}$. In this case, with probability at least $1 - 2\varepsilon_1$

$$\bigl|\log(\alpha) - \log(\widehat{\alpha})\bigr| \leq \frac{\zeta}{2}. \tag{4.6}$$

As a result, with probability at least $1 - 2\varepsilon_1$,

$$\mathcal{K}(\rho, \pi) = \log\bigl(1 + x^{-1}\bigr).$$

Indeed, whenever equation (4.6) holds, the (random) support of ρ is included in the (fixed) support of π, so that the relative entropy of ρ with respect to π is given by the logarithm of the ratio of the lengths of their supports.

Let us now upper-bound $\psi[\widehat{\alpha}(Y_i - \theta)]$ as a suitable function of ρ.

Lemma 4.4. *For any posterior distribution ρ and any $f \in L_2(\rho)$,*

$$\psi\left[\int \rho(\mathrm{d}\beta) f(\beta)\right] \le \int \rho(\mathrm{d}\beta) \log\left[1 + f(\beta) + \frac{1}{2} f(\beta)^2 + \frac{a}{2} \mathbb{V}\mathrm{ar}_\rho(f)\right],$$

where $a \le 4.43$ is the numerical constant of equation (7.1), page 1174.

Applying this inequality to $f(\beta) = \beta(Y_i - \theta)$ in the first place and $f(\beta) = \beta(\theta - Y_i)$ to get the reversed inequality, we obtain

$$\psi\big[\widehat{\alpha}(Y_i - \theta)\big] \le \int \rho(\mathrm{d}\beta) \log\left\{1 + \beta(Y_i - \theta) + \frac{1}{2}\left[\beta^2 + \frac{a}{3} x^2 \alpha^2 \sinh(\zeta/2)^2\right](Y_i - \theta)^2\right\}, \tag{4.7}$$

$$\psi\big[\widehat{\alpha}(\theta - Y_i)\big] \le \int \rho(\mathrm{d}\beta) \log\left\{1 + \beta(\theta - Y_i) + \frac{1}{2}\left[\beta^2 + \frac{a}{3} x^2 \alpha^2 \sinh(\zeta/2)^2\right](Y_i - \theta)^2\right\}. \tag{4.8}$$

Let us now recall the fundamental PAC-Bayes inequality, concerned with any function $(\beta, y) \mapsto f(\beta, y) \in \mathbb{L}_1(\pi \otimes \mathbb{P})$, where $\mathbb{P}$ is the distribution of Y, such that $\inf f > -1$.

$$\begin{aligned}
&\mathbb{E}\left\{\exp\left(\int \rho(\mathrm{d}\beta) \sum_{i=1}^n \log\big[1 + f(\beta, Y_i)\big] - n \log\big[1 + \mathbb{E}\big(f(\beta, Y)\big)\big] - \mathcal{K}(\rho, \pi)\right)\right\} \\
&\quad \le \mathbb{E}\left\{\int \pi(\mathrm{d}\beta) \exp\left(\sum_{i=1}^n \log\big[1 + f(\beta, Y_i)\big] - n \log\big[1 + \mathbb{E}\big(f(\beta, Y)\big)\big]\right)\right\} = 1,
\end{aligned}$$

according to [7], p. 159, and Fubini's lemma for the last equality. Thus, from Chebyshev's inequality, with probability at least $1 - \varepsilon_2$,

$$\begin{aligned}
&\int \rho(\mathrm{d}\beta) \sum_{i=1}^n \log\big[1 + f(\beta, Y_i)\big] \\
&\quad < n \int \rho(\mathrm{d}\beta) \log\big[1 + \mathbb{E}\big(f(\beta, Y)\big)\big] + \mathcal{K}(\rho, \pi) + \log\big(\varepsilon_2^{-1}\big) \\
&\quad \le n \int \rho(\mathrm{d}\beta) \mathbb{E}\big(f(\beta, Y)\big) + \mathcal{K}(\rho, \pi) + \log\big(\varepsilon_2^{-1}\big).
\end{aligned}$$

Applying this inequality to the case we are interested in, that is to equations (4.7) and (4.8), we obtain with probability at least $1 - 2\varepsilon_1 - 2\varepsilon_2$ that

$$\begin{aligned}
\widehat{\alpha} r(\theta_+) < \widehat{\alpha}(m - \theta_+) &+ \frac{1}{2}\left[\widehat{\alpha}^2 + \frac{(a+1)}{3} x^2 \alpha^2 \sinh(\zeta/2)^2\right]\big[v + (m - \theta_+)^2\big] \\
&+ \frac{\log(1 + x^{-1}) + \log(\varepsilon_2^{-1})}{n}
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\alpha} r(\theta_-) > \widehat{\alpha}(m - \theta_-) &- \frac{1}{2}\left[\widehat{\alpha}^2 + \frac{(a+1)}{3} x^2 \alpha^2 \sinh(\zeta/2)^2\right]\big[v + (m - \theta_-)^2\big] \\
&- \frac{\log(1 + x^{-1}) + \log(\varepsilon_2^{-1})}{n}.
\end{aligned}$$

Let us put $\theta_+ - m = m - \theta_- = \sqrt{\gamma v}$ and let us look for some value of γ ensuring that $r(\theta_+) < 0 < r(\theta_-)$, implying that $\theta_- \le \widehat{\theta} \le \theta_+$. Let us choose

$$\begin{aligned}\alpha &= \sqrt{\frac{2[\log(1+x^{-1}) + \log(\varepsilon_2^{-1})]}{n[1-((a+1)/3)x^2\sinh(\zeta/2)^2](1+\gamma)v}},\\ \widehat{\alpha} &= \sqrt{\frac{2[\log(1+x^{-1}) + \log(\varepsilon_2^{-1})]}{n[1-((a+1)/3)x^2\sinh(\zeta/2)^2](1+\gamma)\widehat{v}}} = \alpha\sqrt{\frac{v}{\widehat{v}}},\end{aligned} \tag{4.9}$$

assuming that x will be chosen later on such that

$$\frac{(a+1)}{3}x^2\sinh(\zeta/2)^2 < 1.$$

Since

$$\frac{\log(1+x^{-1}) + \log(\varepsilon_2^{-1})}{n} = \frac{\alpha^2}{2}\left(1 - \frac{(a+1)}{3}x^2\sinh(\zeta/2)^2\right)(1+\gamma)v,$$

we obtain with probability at least $1 - 2\varepsilon_1 - 2\varepsilon_2$

$$r(\theta_+) < -\sqrt{\gamma v} + \frac{\alpha v(1+\gamma)}{2}\left(\frac{\widehat{\alpha}}{\alpha} + \frac{\alpha}{\widehat{\alpha}}\right) \le -\sqrt{\gamma v} + \alpha v(1+\gamma)\cosh(\zeta/2)$$

and

$$r(\theta_-) > \sqrt{\gamma v} - \frac{\alpha v(1+\gamma)}{2}\left(\frac{\widehat{\alpha}}{\alpha} + \frac{\alpha}{\widehat{\alpha}}\right) \ge \sqrt{\gamma v} - \alpha v(1+\gamma)\cosh(\zeta/2).$$

Therefore, if we choose γ such that $\sqrt{\gamma v} = \alpha v(1+\gamma)\cosh(\zeta/2)$, we obtain with probability at least $1 - 2\varepsilon_1 - 2\varepsilon_2$ that $r(\theta_+) < 0 < r(\theta_-)$, and therefore that $\theta_- < \widehat{\theta} < \theta_+$. The corresponding value of γ is $\gamma = \eta/(1-\eta)$, where

$$\eta = \frac{2\cosh(\zeta/2)^2[\log(1+x^{-1}) + \log(\varepsilon_2^{-1})]}{n[1-((a+1)/3)x^2\sinh(\zeta/2)^2]}.$$

Proposition 4.5. *With probability at least* $1 - 2\varepsilon_1 - 2\varepsilon_2$, *the estimator* $\widehat{\theta}_{\widehat{\alpha}}$ *defined by equation* (4.5), *page* 1161, *where* $\widehat{\alpha}$ *is set as in* (4.9), *page* 1165, *satisfies*

$$|\widehat{\theta}_{\widehat{\alpha}} - m| \le \sqrt{\frac{\eta v}{1-\eta}} \le \sqrt{\frac{\eta\widehat{v}}{1-\eta}}\exp(\zeta/2) \le \sqrt{\frac{\eta v}{1-\eta}}\exp(\zeta).$$

The optimal value of x is the one minimizing

$$\frac{\log(1+x^{-1}) + \log(\varepsilon_2^{-1})}{1-((a+1)/3)x^2\sinh(\zeta/2)^2}.$$

Assuming ζ to be small, the optimal x will be large, so that $\log(1+x^{-1}) \simeq x^{-1}$, and we can choose the approximately optimal value

$$x = \left[\frac{2(a+1)}{3}\log\left(\varepsilon_2^{-1}\right)\right]^{-1/3}\sinh(\zeta/2)^{-2/3}.$$

Let us discuss now the question of balancing ε_1 and ε_2. Let us put $\varepsilon = \varepsilon_1 + \varepsilon_2$ and let $y = \varepsilon_1/\varepsilon$. Optimizing y for a fixed value of ε could be done numerically, although it seems difficult to obtain a closed formula. However, the

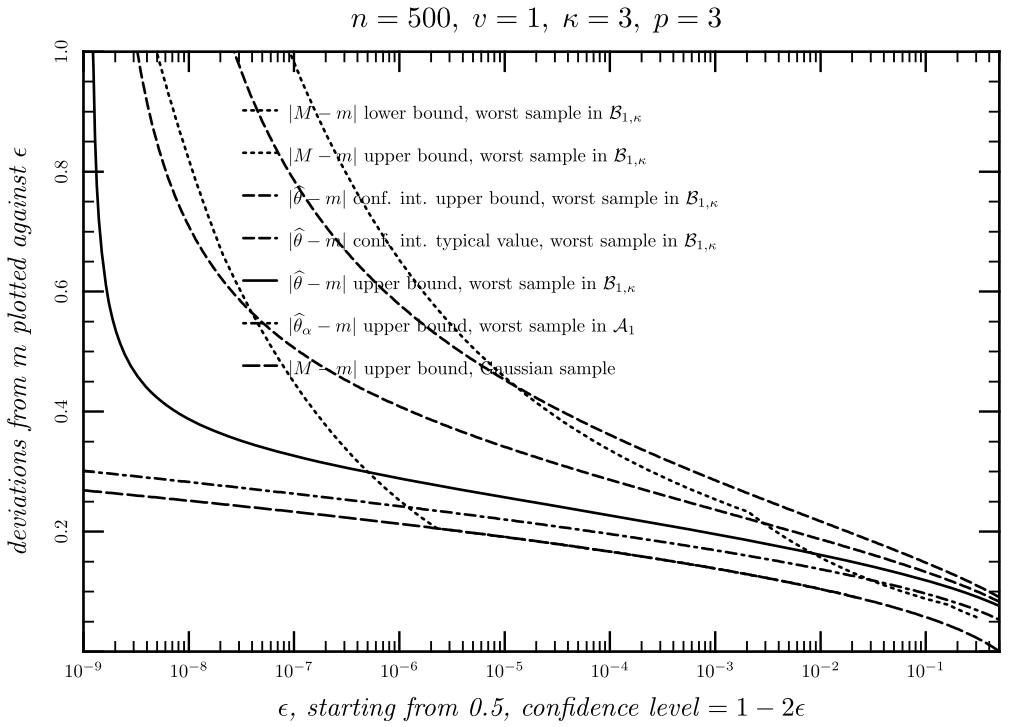

Fig. 9. Comparison of mean estimates, for a sample with bounded kurtosis and unknown variance.

entropy term in η can be written as $\log(1+x^{-1}) + \log(\varepsilon^{-1}) - \log(1-y)$. Since ζ decreases, and therefore the almost optimal x above increases, when y increases, we will get an optimal order of magnitude (up to some constant less than 2) for the bound if we balance $-\log(1-y)$ and $\log(1+x^{-1})$, resulting in the choice $y = (1+x)^{-1}$, where x is approximately optimized as stated above (this choice of x depends on y, so we end up with an equation for x, which can be solved using an iterative approach). This results, with probability at least $1-2\varepsilon$, in an upper bound for $|\widehat{\theta} - m|$ equivalent for large values of the sample size n to

$$\sqrt{\frac{2\log(\varepsilon^{-1})v}{n}}.$$

Thus we recover, as desired, the same asymptotics as when the variance is known.

Let us illustrate what we get when $n = 500$ or $n = 1000$ and it is known that $\kappa \leq 3$ (Figs 9 and 10).

On these figures, we have plotted upper and lower bounds for the deviations of the empirical mean when the sample distribution[3] is the least favourable one in $\mathcal{B}_{1,\kappa}$. (These bounds will be proved later on. The upper bound is computed by taking the minimum of three bounds, explaining the discontinuities of its derivative.) What we see on the $n = 500$ example is that our bound remains of the same order as the Gaussian bound up to confidence levels of order $1 - 10^{-8}$, whereas this is not the case with the empirical mean.

In the case when $n = 1000$, we see that our estimator possibly improves on the empirical mean in the range of confidence levels going from $1 - 10^{-2}$ to $1 - 10^{-6}$ and is a proved winner in the range going from $1 - 10^{-6}$ to $1 - 10^{-14}$.

Let us see now the influence of κ and plot the curves corresponding to increasing values of n and κ (Figs 11–13). When we double the kurtosis letting $n = 1000$, we follow the Gaussian curve up to confidence levels of $1 - 10^{-10}$ instead of $1 - 10^{-14}$. This is somehow the maximum kurtosis for which the bounds are satisfactory for this sample size. Looking at Proposition 4.2 (page 1161), we see that the bound in first approximation depends on the ratio $\chi/n = \kappa/n$

[3] $\mathcal{B}_{1,\kappa}$ is defined by (1.2), page 1149.

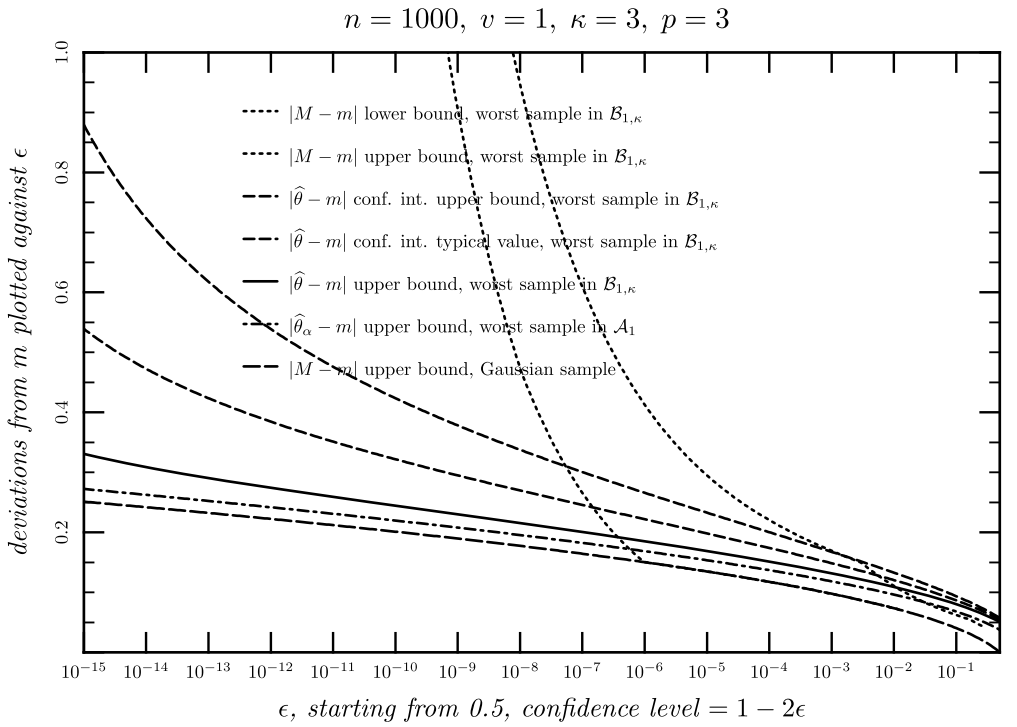

Fig. 10. Comparison of mean estimates, for a larger sample size.

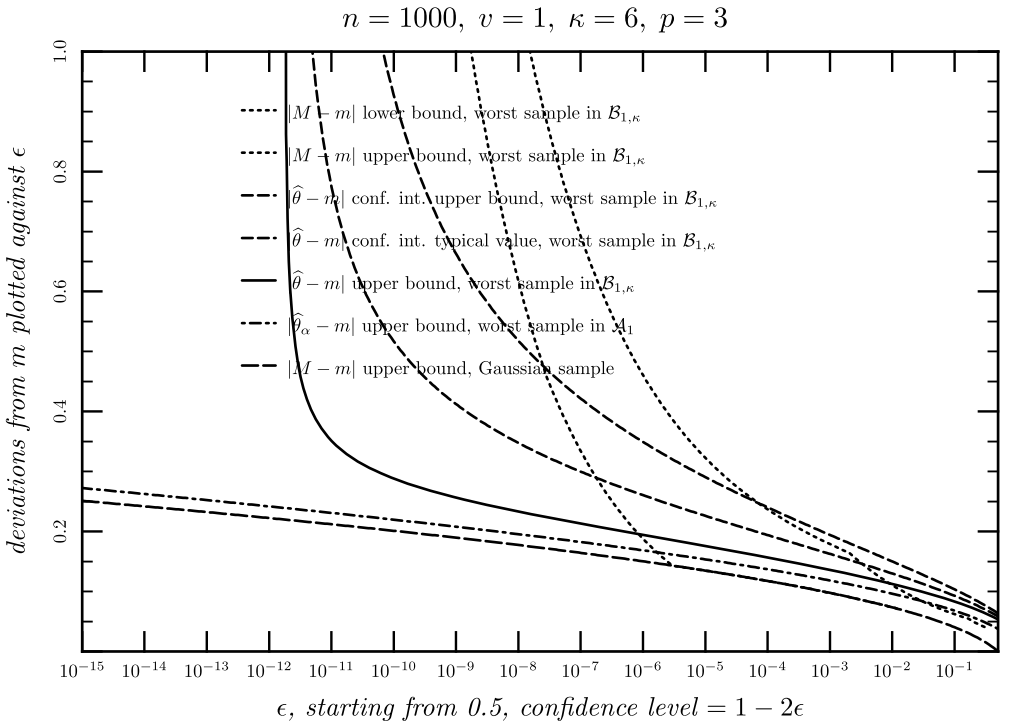

Fig. 11. Comparison of mean estimates, for a sample distribution with larger kurtosis.

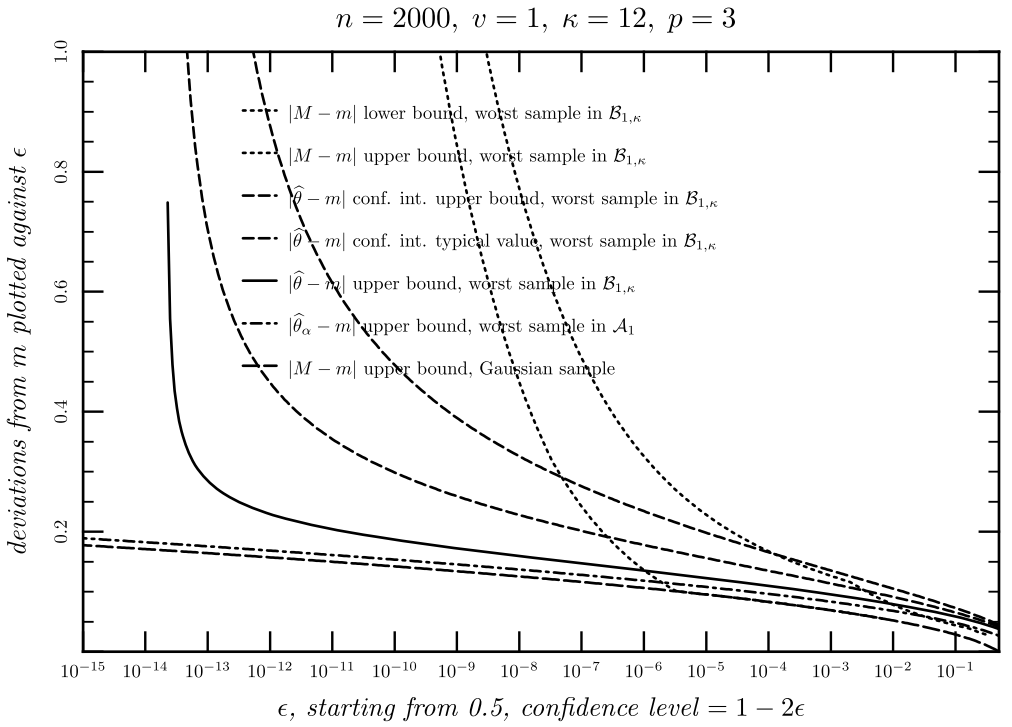

Fig. 12. Comparison of mean estimates, when the kurtosis is even larger.

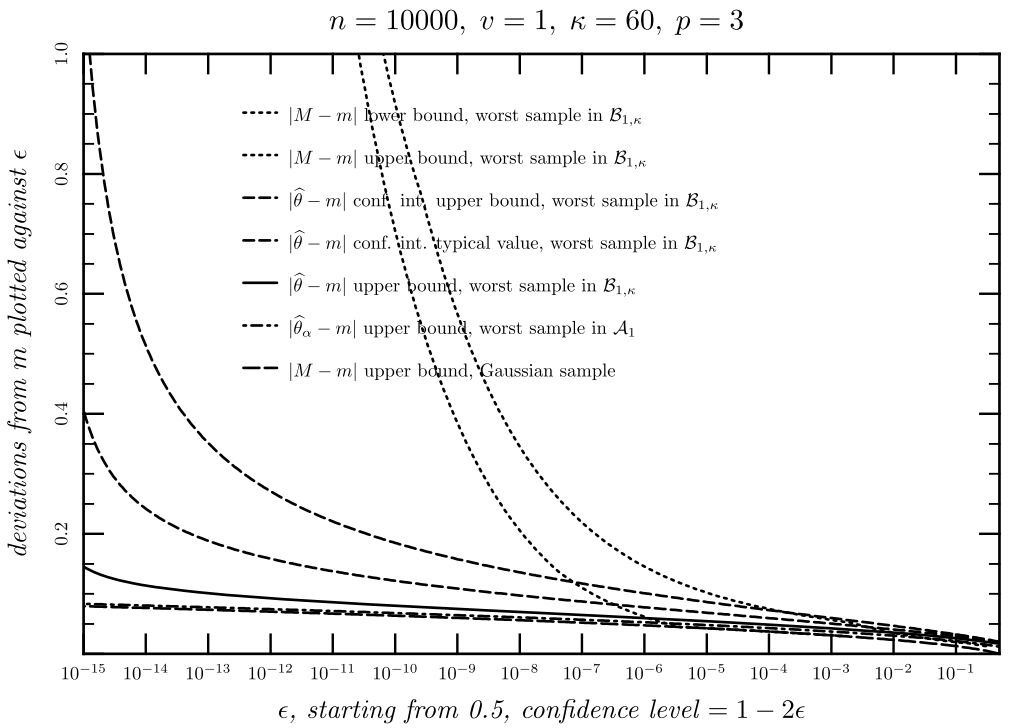

Fig. 13. Comparison of mean estimates, for a very large kurtosis and accordingly large sample size.

(when $p = 3$), suggesting that to obtain similar performances, we have to take n proportional to κ, which gives a minimum sample size of $n = 1000\kappa/6$, if we want to follow the Gaussian curve up to confidence levels of order at least $1 - 10^{-10}$.

These curves suggest another approach to choose the kurtosis parameter κ. It is to use the largest value of the kurtosis with a low impact on the bound of Proposition 4.5 (page 1165), given the sample size. This leads, when in doubt about the true kurtosis value, for sample sizes $n \geq 1000$, to set, according to the previous rule of thumb, the kurtosis in the definition of the estimator to the value $\kappa_{\max} = 6n/1000$. Doing so, we get almost the same deviations as if the sample distribution were Gaussian, at levels of confidence up to $1 - 10^{-10}$, for the largest range of (possibly non-Gaussian) sample distributions.

5. Upper bounds for the deviations of the empirical mean

In the previous sections, we compared new mean estimators with the empirical mean. We will devote the end of this paper to prove the bounds on the empirical mean used in these comparisons.

This section deals with upper bounds, whereas the next one will study corresponding lower bounds.

Let us start with the case when the sample distribution may be any probability measure with a finite variance. It is natural in this situation to bound the deviations of the empirical mean

$$M = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

applying Chebyshev's inequality to its second moment, to conclude that

$$\mathbb{P}\left(|M - m| \geq \sqrt{\frac{v}{2\varepsilon n}}\right) \leq 2\varepsilon. \tag{5.1}$$

This is in general close to optimal, as will be shown later when we will compute corresponding lower bounds.

When the sample distribution has a finite kurtosis, it is possible to take this into account to refine the bound. The analysis becomes less straightforward, and will be carried out in this section. The following bound uses a truncation argument, allowing to study separately the behaviour of small and large values. It is to our knowledge a new result. We will show later in this paper that its leading term is essentially tight – up to a factor $(\frac{\kappa}{\kappa-1})^{1/4}$ – when the proper asymptotic is considered.

Proposition 5.1. *For any probability distribution whose kurtosis is not greater than κ, the empirical mean M is such that with probability at least $1 - 2\varepsilon$,*

$$\begin{aligned}\frac{|M-m|}{\sqrt{v}} &\leq \inf_{\lambda\in(0,1)} \sqrt{\frac{2\log(\lambda^{-1}\varepsilon^{-1})}{n}} + \frac{\sqrt{\kappa}\log(\lambda^{-1}\varepsilon^{-1})}{3n} \\ &\quad + \left(\frac{\kappa}{2(1-\lambda)n^3\varepsilon}\right)^{1/4}\left[1 + \frac{3(n-1)\kappa\log(\lambda^{-1}\varepsilon^{-1})^2}{4^3(1+\sqrt{2})^4 n^2}\right]^{1/4} \\ &\underset{\substack{n\varepsilon\to 0\\ \varepsilon^{1/n}\to 1}}{\simeq} \left(\frac{\kappa}{2n^3\varepsilon}\right)^{1/4}.\end{aligned}$$

Instead of minimizing the bound in λ, one can also take for simplicity

$$\lambda = \min\left\{\frac{1}{2}, 2^{7/4}\left(\frac{n\varepsilon}{\kappa}\right)^{1/4}\sqrt{\log\left(\frac{\kappa}{2n\varepsilon^5}\right)}\right\}.$$

We see that there are two regimes in the behaviour of the deviations of M. A Gaussian regime for levels of confidence less than $1 - 1/n$ and long tail regime for higher confidence levels, depending on the value of the kurtosis κ.

In addition to this, let us also put forward the fact that, even in the simple case when the mean m is known, estimating the variance under a kurtosis hypothesis at high confidence levels cannot be done using the empirical estimator

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2.$$

Indeed, assuming without loss of generality that $m = 0$ and computing the quadratic mean

$$\mathbb{E}\left\{\left[M_2 - \mathbb{E}\left(Y^2\right)\right]^2\right\} = \frac{\mathbb{E}(Y^4) - \mathbb{E}\left(Y^2\right)^2}{n} = \frac{(\kappa-1)}{n}\mathbb{E}\left(Y^2\right)^2,$$

we can only conclude, using Chebyshev's inequality, that with probability at least $1 - 2\varepsilon$

$$\mathbb{E}\left(Y^2\right) \le \frac{M_2}{1 - \sqrt{(\kappa-1)/(2n\varepsilon)}},$$

a bound which blows up at level of confidence $\varepsilon = \frac{\kappa-1}{2n}$, and which we do not suspect to be substantially improvable in the worst case. In contrast to this, Propositions 4.1 and 4.2 (page 1161) provide variance estimators with high confidence levels.

6. Lower bounds

6.1. *Lower bound for Gaussian sample distributions*

This lower bound is well known. We recall it here for the sake of completeness.

The empirical mean has optimal deviations when the sample distribution is Gaussian in the following precise sense.

Proposition 6.1. *For any estimator of the mean $\widehat{\theta}:\mathbb{R}^n \to \mathbb{R}$, any variance value $v > 0$, and any deviation level $\eta > 0$, there is some Gaussian measure $\mathcal{N}(m, v)$ (with variance v and mean m) such that the i.i.d. sample of length n drawn from this distribution is such that*

$$\mathbb{P}(\widehat{\theta} \ge m + \eta) \ge \mathbb{P}(M \ge m + \eta) \quad \text{or} \quad \mathbb{P}(\widehat{\theta} \le m - \eta) \ge \mathbb{P}(M \le m - \eta),$$

where $M = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the empirical mean.

This means that any distribution free symmetric confidence interval based on the (supposedly known) value of the variance has to include the confidence interval for the empirical mean of a Gaussian distribution, whose length is exactly known and equal to the properly scaled quantile of the Gaussian measure.

Let us state this more precisely. With the notations of the previous proposition

$$\mathbb{P}(M \ge m + \eta) = \mathbb{P}(M \le m - \eta) = G\left[\left(\sqrt{\frac{n}{v}}\eta, +\infty\right)\right] = 1 - F\left(\sqrt{\frac{n}{v}}\eta\right),$$

where G is the standard normal measure and F its distribution function.

The upper bounds proved in this paper can be decomposed into

$$\mathbb{P}(\widehat{\theta} \ge m + \eta) \le \varepsilon \quad \text{and} \quad \mathbb{P}(\widehat{\theta} \le m - \eta) \le \varepsilon,$$

although we preferred for simplicity to state them in the slightly weaker form $\mathbb{P}(|\theta - m| \ge \eta) \le 2\varepsilon$.

As the Gaussian shift model made of Gaussian sample distributions with a given variance and varying means, is included in all the models we are considering in this paper, we necessarily should have according to the previous proposition

$$\varepsilon \ge 1 - F\left(\sqrt{\frac{n}{v}}\eta\right),$$

which can be also written as

$$\eta \geq \sqrt{\frac{v}{n}} F^{-1}(1-\varepsilon).$$

Therefore some visualization of the quality of our bounds can be obtained by plotting $\varepsilon \mapsto \eta$ against $\varepsilon \mapsto \sqrt{\frac{v}{n}} F^{-1}(1-\varepsilon)$, as we did in the previous sections.

Let us remark eventually that the assumed symmetry of the confidence region is not a real limitation. Indeed, if we can prove for any given estimator $\widehat{\theta}$ that for any Gaussian sample distribution with a given variance v,

$$\mathbb{P}\big[m \geq \widehat{\theta} + \eta_+(\varepsilon)\big] \leq \varepsilon \quad \text{and} \quad \mathbb{P}\big[m \leq \widehat{\theta} - \eta_-(\varepsilon)\big] \leq \varepsilon,$$

then we may consider for any value of ε the estimator with symmetric confidence levels defined as

$$\widehat{\theta}_s = \widehat{\theta} + \frac{\eta_+(\varepsilon) - \eta_-(\varepsilon)}{2}.$$

This symmetric estimator is such that for any Gaussian sample distribution with variance v,

$$\mathbb{P}\left[m \geq \widehat{\theta}_s + \frac{\eta_-(\varepsilon) + \eta_+(\varepsilon)}{2}\right] \leq \varepsilon,$$

$$\mathbb{P}\left[m \leq \widehat{\theta}_s - \frac{\eta_-(\varepsilon) + \eta_+(\varepsilon)}{2}\right] \leq \varepsilon.$$

Thus, applying the previous proposition, we obtain that

$$\frac{\eta_+(\varepsilon) + \eta_-(\varepsilon)}{2} \geq \sqrt{\frac{v}{n}} F^{-1}(1-\varepsilon).$$

6.2. *Lower bound for the deviations of the empirical mean depending on the variance*

In the following proposition, we state a lower bound for the deviations of the empirical mean when the sample distribution[4] is the least favourable in $\mathcal{A}_{v_{\max}}$ (meaning the distribution for which the deviations of the empirical mean are the largest).

Proposition 6.2. *For any value of the variance v, any deviation level $\eta > 0$, there is some distribution with variance v and mean 0 such that the i.i.d. sample of size n drawn from it satisfies*

$$\mathbb{P}(M \geq \eta) = \mathbb{P}(M \leq -\eta) \geq \frac{v(1 - v/(\eta^2 n^2))^{n-1}}{2n\eta^2}.$$

Thus, as soon as $\varepsilon \leq (2\mathrm{e})^{-1}$, with probability at least 2ε,

$$|M - m| \geq \sqrt{\frac{v}{2n\varepsilon}} \left(1 - \frac{2\mathrm{e}\varepsilon}{n}\right)^{(n-1)/2}.$$

Let us remark that this bound is pretty tight, since, according to equation (5.1), page 1169, with probability at least $1 - 2\varepsilon$,

$$|M - m| \leq \sqrt{\frac{v}{2n\varepsilon}}.$$

This can also be observed on the plots following Proposition 2.4 (page 1153).

[4] $\mathcal{A}_{v_{\max}}$ is defined by (1.1), page 1149.

6.3. *Lower bound for the deviations of empirical mean depending on the variance and the kurtosis*

Let us now refine the previous lower bound by taking into account the kurtosis κ of the sample distribution, assuming of course that it is finite.

Proposition 6.3. *As soon as $\varepsilon^{-1} \geq n \geq 16$, there exists a sample distribution with mean m, finite variance v and finite kurtosis κ, such that with probability at least 2ε,*

$$|M-m| \geq \max\left\{\left[\frac{(\kappa-1)(1-8\varepsilon)}{4n\varepsilon}\right]^{1/4}, \left[\frac{(\kappa-1)}{2n\varepsilon}\left[1-\left(\frac{n\varepsilon}{16}\right)^{1/4}-4\varepsilon\right]\right]^{1/4} - \sqrt{\frac{\log[16/(n\varepsilon)]v}{2n}}\right\}\sqrt{\frac{v}{n}}.$$

Let us remark that the asymptotic behaviour of this lower bound when $n\varepsilon$ and $\log(\varepsilon^{-1})/n$ both tend to zero matches the upper bound of Proposition 5.1 (page 1169) up to a multiplicative factor $(\frac{\kappa}{\kappa-1})^{1/4} \leq 1.11$ when the kurtosis is $\kappa \geq 3$, which is the kurtosis of the Gaussian distribution.

The plots following Proposition 4.5 (page 1165) show that this lower bound is not too far from the upper bound obtained by combining Propositions 5.1 (page 1169) and Proposition 7.3 (page 1176) and equation (5.1), page 1169.

7. Proofs

7.1. *Proof of Corollary 4.3 (page 1161)*

Let us remark first that condition (4.4), page 1161, implies condition (4.3), page 1160, as can be easily checked. Putting

$$x = \sqrt{\frac{n}{(\kappa-1)[4\log(\varepsilon_1^{-1})+1/2]}},$$

so that $p = \lfloor x \rfloor$, we can also check that $p-1 \geq x/2$ and $n-p+1 \geq n/2$. We can then write

$$\begin{aligned}
&\sqrt{\frac{2\chi\log(\varepsilon_1^{-1})}{n-r}}\left(1+\frac{4\log(\varepsilon_1^{-1})}{q}\right)\\
&\quad\leq \sqrt{\frac{2(\kappa-1)\log(\varepsilon_1^{-1})}{n}}\exp\left(\frac{1}{(\kappa-1)(p-1)}+\frac{p-1}{2(n-p+1)}+\frac{4\log(\varepsilon_1^{-1})p}{n-p+1}\right)\\
&\quad\leq \sqrt{\frac{2(\kappa-1)\log(\varepsilon_1^{-1})}{n}}\exp\left(\frac{2}{(\kappa-1)x}+\frac{2x[4\log(\varepsilon_1^{-1})+1/2]}{n}\right)\\
&\quad= \sqrt{\frac{2(\kappa-1)\log(\varepsilon_1^{-1})}{n}}\exp\left(4\sqrt{\frac{4\log(\varepsilon_1^{-1})+1/2}{(\kappa-1)n}}\right).
\end{aligned}$$

7.2. *Proof of Lemma 4.4 (page 1164)*

Let us introduce some modification of ψ in order to improve the compromise between $\inf\psi''$ and $\sup\psi$. Let us put $\overline{\psi}(x) = \log(1+x+x^2/2)$. We would like to squeeze some function χ between ψ and $\overline{\psi}$, in such a way that $\inf\chi'' = \inf\overline{\psi}''$. This will be better than using ψ itself since

$$\inf\overline{\psi}'' = -1/4, \quad \text{whereas } \inf\psi'' = -2.$$

Indeed these two values can be computed in the following way. Let us put $\varphi(x) = \exp[\overline{\psi}(x)] = 1+x+x^2/2$. It is easy to check that

$$\overline{\psi}''(x) = -\varphi(x)^{-1}\big[1-\varphi(x)^{-1}\big], \qquad \psi''(x) = \varphi(-x)^{-1}\big[1-\varphi(-x)^{-1}\big],$$

implying that $\overline{\psi}''(x) \geq -1/4$. This inequality becomes an equality when $\varphi(x) = 2$, that is when $x = \sqrt{3} - 1 \simeq 0.73$. In the same way $\psi''(x) \geq -2$ and equality is reached when $\varphi(-x) = 1/2$, that is when $x = 1$. We are going to build a function χ which follows ψ when $x \leq x_1$, where x_1 satisfies $\psi''(x_1) = -1/4$. The value of x_1 is computed from the equation $\varphi(-x_1)^{-1} = (1 + \sqrt{2})/2$. Let $y_1 = \psi(x_1)$ and $p_1 = \psi'(x_1)$. They have the following values

$$x_1 = 1 - \sqrt{4\sqrt{2} - 5} \simeq 0.1895,$$
$$y_1 = -\log\bigl[2\bigl(\sqrt{2} - 1\bigr)\bigr] \simeq 0.1882,$$
$$p_1 = \frac{\sqrt{4\sqrt{2} - 5}}{2(\sqrt{2} - 1)} \simeq 0.978.$$

After x_1, we continue χ with a quadratic function, until its derivatives cancels. Thus, the second derivative of χ being less than the second derivative of $\overline{\psi}$ at each point of the positive real line, we are sure that $\chi(x) \leq \overline{\psi}(x)$ for any $x \in \mathbb{R}$. The function χ built in this way satisfies the equation

$$\chi(x) = \begin{cases} \psi(x), & x \leq x_1, \\ y_1 + p_1(x - x_1) - \frac{1}{8}(x - x_1)^2, & x_1 \leq x \leq x_1 + 4p_1, \\ y_1 + 2p_1^2 \leq 2.103, & x \geq x_1 + 4p_1. \end{cases}$$

As we have proved, and as can be seen in Fig. 14, the function χ is such that

$$\psi(x) \leq \chi(x) \leq \overline{\psi}(x), \quad x \in \mathbb{R}.$$

Let us now compare χ with a suitable convex function (in order to apply Jensen's inequality). Let us introduce to this purpose the function $\overline{\chi}_{x_*} = \chi(x) + \frac{1}{8}(x - x_*)^2$, which is convex for any choice of the parameter $x_* \in \mathbb{R}$.

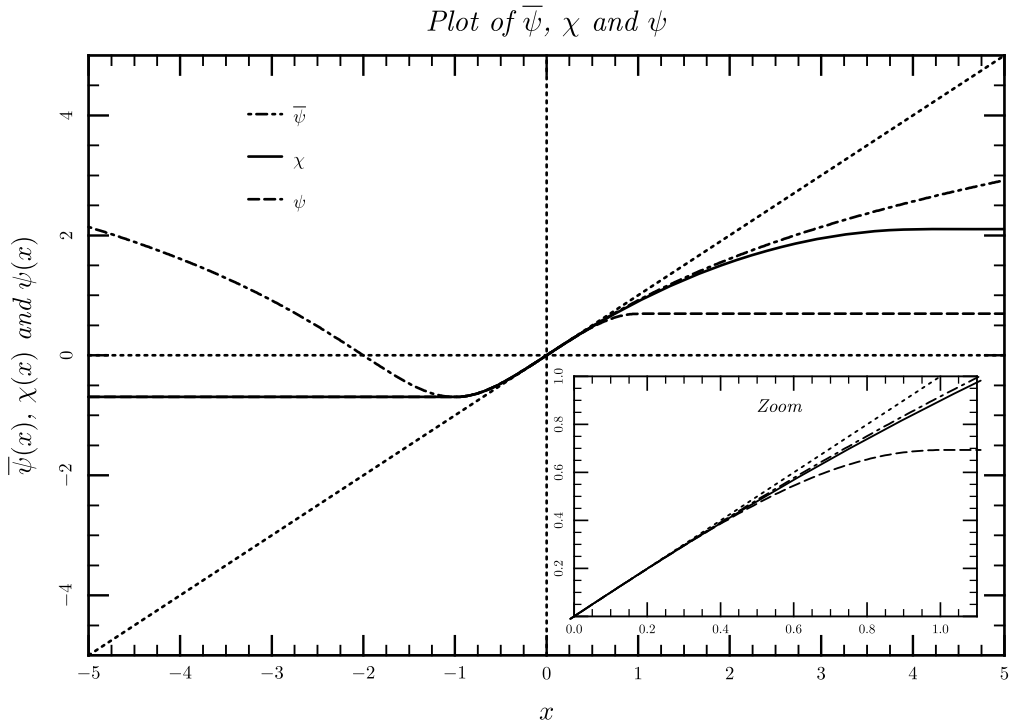

Fig. 14. Plot of the modified influence functions used in the proof.

Let us consider as in the lemma we have to prove some function $f \in L_2(\rho)$ and let us choose $x_* = \int \rho(\mathrm{d}\beta) f(\beta)$ and put $\int \rho(\mathrm{d}\beta)[f(\beta) - \int \rho f]^2 = \mathbb{V}\mathrm{ar}_\rho(f)$. We obtain by Jensen's inequality

$$\begin{aligned}\psi(x_*) \le \chi(x_*) = \overline{\chi}_{x_*}(x_*) &= \overline{\chi}_{x_*}\left[\int \rho(\mathrm{d}\beta) f(\beta)\right] \\ &\le \int \rho(\mathrm{d}\beta)\overline{\chi}_{x_*}\big[f(\beta)\big] = \int \rho(\mathrm{d}\beta)\chi\big[f(\beta)\big] + \frac{1}{8}\,\mathbb{V}\mathrm{ar}_\rho(f).\end{aligned}$$

On the other hand, it is clear that

$$\psi(x_*) \le \int \rho(\mathrm{d}\beta)\chi\big[f(\beta)\big] - \inf\chi + \sup\psi = \int \rho(\mathrm{d}\beta)\chi\big[f(\beta)\big] + \log(4).$$

We have proved

Lemma 7.1. *For any posterior distribution ρ and any $f \in L_2(\rho)$,*

$$\psi\left[\int \rho(\mathrm{d}\beta) f(\beta)\right] \le \int \rho(\mathrm{d}\beta)\chi\big[f(\beta)\big] + \min\left\{\log(4), \frac{1}{8}\,\mathbb{V}\mathrm{ar}_\rho(f)\right\}.$$

To end the proof of Lemma 4.4 (page 1164), it remains to establish that for any $x \in \mathbb{R}$ and $y \in \mathbb{R}_+$,

$$\chi(x) + \min\left\{\log(4), \frac{y}{8}\right\} \le \log\left[1 + x + x^2/2 + \frac{ay}{2}\right],$$

where

$$a = \frac{3\exp[\sup(\chi)]}{4\log(4)} = \frac{3\exp(y_1 + 2p_1^2)}{4\log(4)} \le 4.43. \tag{7.1}$$

Indeed a should satisfy

$$a \ge \frac{2}{y}\left[\exp\big[\chi(x)\big]\min\left\{4, \exp\left(\frac{y}{8}\right)\right\} - \left(1 + x + \frac{x^2}{2}\right)\right], \quad x \in \mathbb{R}, y \in \mathbb{R}_+.$$

Since $\exp[\chi(x)] \le 1 + x + \frac{x^2}{2}$, the right-hand side of this inequality is less than

$$\frac{2\exp[\chi(x)]}{y}\left(\min\left\{4, \exp\left(\frac{y}{8}\right)\right\} - 1\right).$$

As $y \mapsto y^{-1}[\exp(\frac{y}{8}) - 1]$ is increasing on $\mathbb{R}_+$, this last expression reaches its maximum when $x \in \arg\max\chi$ and $\exp(\frac{y}{8}) = 4$, and is then equal to $\frac{3\exp[\sup(\chi)]}{4\log(4)}$, which is the value stated for a in equation (7.1) above.

7.3. *Proof of Proposition* 5.1 (*page* 1169)

Consider the function

$$g(x) = \begin{cases} x - \log\big(1 + x + x^2/2\big), & x \ge 0, \\ x + \log\big(1 - x + x^2/2\big), & x \le 0. \end{cases}$$

This function could also be defined as $g(x) = x - \psi(x)$, where ψ is the wide version of the influence function defined by equation (2.2), page 1151. It is such that

$$g'(x) = \frac{x^2}{2(1 + x + x^2/2)} \le \frac{x^2}{2}, \quad x \ge 0.$$

Therefore

$$0 \le g(x) \le \frac{x^3}{6}, \quad x \ge 0,$$

implying by symmetry that

$$\bigl|g(x)\bigr| \le \frac{|x|^3}{6}, \quad x \in \mathbb{R}.$$

We can also remark that

$$g'(x) \le \frac{x}{2(1+\sqrt{2})}, \quad x \ge 0,$$

implying that

$$\bigl|g(x)\bigr| \le \frac{x^2}{4(1+\sqrt{2})}, \quad x \in \mathbb{R}.$$

As it is also obvious that $|g(x)| \le |x|$, we get

Lemma 7.2.

$$\bigl|g(x)\bigr| \le \min\left\{|x|, \frac{x^2}{4(1+\sqrt{2})}, \frac{|x|^3}{6}\right\}.$$

Now let us write

$$M = m + \frac{1}{\alpha n}\sum_{i=1}^{n} \psi\bigl[\alpha(Y_i - m)\bigr] + \frac{1}{\alpha n}\sum_{i=1}^{n} G_i,$$

where $G_i = g[\alpha(Y_i - m)]$ and ψ is the wide influence function of equation (2.2), page 1151. As we have already seen in Proposition 2.2 (page 1152), with probability at least $1 - 2\varepsilon_1$,

$$\left|\frac{1}{n\alpha}\sum_{i=1}^{n} \psi\bigl[\alpha(Y_i - m)\bigr]\right| \le \frac{\alpha v}{2} + \frac{\log(\varepsilon_1^{-1})}{n\alpha}.$$

On the other hand with probability at least $1 - \varepsilon_2$

$$\left|\frac{1}{n\alpha}\sum_{i=1}^{n} G_i\right| \le \frac{1}{n\alpha}\sum_{i=1}^{n} |G_i| \le \frac{\mathbb{E}(|G|)}{\alpha} + \frac{\mathbb{E}[(\sum_{i=1}^{n} |G_i| - \mathbb{E}(|G|))^4]^{1/4}}{n\alpha\varepsilon_2^{1/4}}.$$

Let us put $H_i = |G_i|$ and let us compute

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n} H_i - \mathbb{E}(H)\right)^4\right] &= \mathbb{E}\bigl\{\bigl[H - \mathbb{E}(H)\bigr]^4\bigr\} + 3(n-1)\mathbb{E}\bigl\{\bigl[H - \mathbb{E}(H)\bigr]^2\bigr\}^2 \\
&= \mathbb{E}\bigl(H^4\bigr) - 4\mathbb{E}\bigl(H^3\bigr)\mathbb{E}(H) + 6\mathbb{E}\bigl(H^2\bigr)\mathbb{E}(H)^2 - 3\mathbb{E}(H)^4 \\
&\quad + 3(n-1)\bigl[\mathbb{E}\bigl(H^2\bigr)^2 - 2\mathbb{E}\bigl(H^2\bigr)\mathbb{E}(H)^2 + \mathbb{E}(H)^4\bigr] \\
&\le \mathbb{E}\bigl(H^4\bigr) + 2\mathbb{E}\bigl(H^2\bigr)\mathbb{E}(H)^2 - 3\mathbb{E}(H)^4 \\
&\quad + 3(n-1)\bigl[\mathbb{E}\bigl(H^2\bigr)^2 - 2\mathbb{E}\bigl(H^2\bigr)\mathbb{E}(H)^2 + \mathbb{E}(H)^4\bigr]
\end{aligned}$$

$$\leq \mathbb{E}\big(H^4\big) + 3(n-1)\mathbb{E}\big(H^2\big)^2 - (3n-2)\mathbb{E}(H)^4$$
$$\leq \alpha^4\kappa v^2 + 3(n-1)\frac{\kappa^2\alpha^8 v^4}{[4(1+\sqrt{2})]^4} - (3n-2)\mathbb{E}(H)^4.$$

Moreover

$$\mathbb{E}(H) \leq \frac{\alpha^3}{6}\mathbb{E}\big(|Y-m|^3\big) \leq \frac{\alpha^3}{6}\sqrt{\kappa v^3}.$$

Thus with probability at least $1-2\varepsilon_1-\varepsilon_2$,

$$|M-m| \leq \frac{\alpha v}{2} + \frac{\log(\varepsilon_1^{-1})}{n\alpha}$$
$$+ \sup_{x\in(0,\sqrt{\kappa\alpha^6 v^3}/6)} \frac{x}{\alpha} + n^{-3/4}\alpha^{-1}\varepsilon_2^{-1/4}\left[\kappa\alpha^4 v^2 + \frac{3(n-1)\kappa^2\alpha^8 v^4}{[4(1+\sqrt{2})]^4} - (3n-2)x^4\right]^{1/4}$$
$$\leq \frac{\alpha v}{2} + \frac{\log(\varepsilon_1^{-1})}{n\alpha} + \frac{\sqrt{\kappa v^3}\alpha^2}{6} + \frac{\sqrt{v}}{n^{3/4}}\left(\frac{\kappa}{\varepsilon_2}\right)^{1/4}\left[1 + \frac{3(n-1)\kappa\alpha^4 v^2}{[4(1+\sqrt{2})]^4}\right]^{1/4}.$$

Let us take

$$\alpha = \sqrt{\frac{2\log(\varepsilon_1^{-1})}{nv}}$$

and let us put $\varepsilon_1 = \lambda\varepsilon$ and $\varepsilon_2 = (1-\lambda)2\varepsilon$.

The bound can either be optimized in λ or we can for simplicity choose λ to balance the following factors

$$\sqrt{\frac{2v\log(\lambda^{-1}\varepsilon^{-1})}{n}} \simeq \frac{\sqrt{v}}{n^{3/4}}\left(\frac{\kappa}{2\varepsilon}\right)^{1/4}\frac{\lambda}{4}.$$

This leads to consider the value

$$\lambda = \min\left\{\frac{1}{2}, 2\sqrt{2\log\left(\frac{\kappa}{2n\varepsilon^5}\right)}\left(\frac{2n\varepsilon}{\kappa}\right)^{1/4}\right\}.$$

As stated in Proposition 5.1 (page 1169), with probability at least $1-2\varepsilon$,

$$|M-m| \leq \sqrt{\frac{2v\log(\lambda^{-1}\varepsilon^{-1})}{n}} + \frac{\sqrt{\kappa v}\log(\lambda^{-1}\varepsilon^{-1})}{3n}$$
$$+ \sqrt{\frac{v}{n}}\left(\frac{\kappa}{2(1-\lambda)n\varepsilon}\right)^{1/4}\left[1 + \frac{3(n-1)\kappa\log(\lambda^{-1}\varepsilon^{-1})^2}{4^3(1+\sqrt{2})^4 n^2}\right]^{1/4}$$
$$\underset{\varepsilon\to 0}{\simeq} \sqrt{\frac{v}{n}}\left(\frac{\kappa}{2n\varepsilon}\right)^{1/4}.$$

Another bound can be obtained applying Chebyshev's inequality directly to the fourth moment of the empirical mean, which however does not reach the right speed when ε is small and n large.

Proposition 7.3. *For any probability distribution whose kurtosis is not greater than κ, the empirical mean M is such that with probability at least $1-2\varepsilon$,*

$$|M-m| \leq \left(\frac{3(n-1)+\kappa}{2n\varepsilon}\right)^{1/4}\sqrt{\frac{v}{n}}.$$

Proof. Let us assume to simplify notations and without loss of generality that $\mathbb{E}(Y) = 0$.

$$\mathbb{E}\big(M^4\big) = \frac{1}{n^4}\sum_{i=1}^{n}\mathbb{E}\big(Y_i^4\big) + \frac{1}{n^4}\sum_{i<j}6\mathbb{E}\big(Y_i^2\big)\mathbb{E}\big(Y_j^2\big) = \frac{\mathbb{E}(Y^4)}{n^3} + \frac{3(n-1)\mathbb{E}(Y^2)^2}{n^3}.$$

It implies that

$$\mathbb{P}\big(|M-m| \geq \eta\big) \leq \frac{\mathbb{E}(M^4)}{\eta^4} \leq \frac{[3(n-1)+\kappa]v^2}{n^3\eta^4},$$

and the result is proved by considering

$$2\varepsilon = \frac{[3(n-1)+\kappa]v^2}{n^3\eta^4}.$$

□

In our comparisons with new estimators, we took the minimum over the three bounds given by Propositions 5.1 (page 1169) and 7.3 (page 1176) and equation (5.1), page 1169.

7.4. *Proof of Proposition 6.1 (page 1170)*

Let us consider the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ of the sample $(Y_i)_{i=1}^n$ obtained when the marginal distributions are respectively the Gaussian measure with variance v and mean $m_1 = -\eta$ and the Gaussian measure with variance v and mean $m_2 = \eta$. We see that, whatever the estimator $\widehat{\theta}$,

$$\begin{aligned}\mathbb{P}_1(\widehat{\theta} \geq m_1+\eta) + \mathbb{P}_2(\widehat{\theta} \leq m_2-\eta) &= \mathbb{P}_1(\widehat{\theta} \geq 0) + \mathbb{P}_2(\widehat{\theta} \leq 0)\\ &\geq (\mathbb{P}_1 \wedge \mathbb{P}_2)(\widehat{\theta} \geq 0) + (\mathbb{P}_1 \wedge \mathbb{P}_2)(\widehat{\theta} \leq 0) \geq |\mathbb{P}_1 \wedge \mathbb{P}_2|,\end{aligned}$$

where $\mathbb{P}_1 \wedge \mathbb{P}_2$ is the measure whose density with respect to the Lebesgue measure (or equivalently with respect to any dominating measure, such as $\mathbb{P}_1 + \mathbb{P}_2$) is the minimum of the densities of $\mathbb{P}_1$ and $\mathbb{P}_2$ and whose total variation is $|\mathbb{P}_1 \wedge \mathbb{P}_2|$.

Now, using the fact that the empirical mean is a sufficient statistics of the Gaussian shift model, it is easy to realize that

$$|\mathbb{P}_1 \wedge \mathbb{P}_2| = \mathbb{P}_1(M \geq m_1+\eta) + \mathbb{P}_2(M \leq m_2-\eta),$$

which obviously proves the proposition.

7.5. *Proof of Proposition 6.2 (page 1171)*

Let us consider the distribution with support $\{-n\eta, 0, n\eta\}$ defined by

$$\mathbb{P}\big(\{n\eta\}\big) = \mathbb{P}\big(\{-n\eta\}\big) = \big[1-\mathbb{P}\big(\{0\}\big)\big]/2 = \frac{v}{2n^2\eta^2}.$$

It satisfies $\mathbb{E}(Y) = 0$, $\mathbb{E}(Y^2) = v$ and

$$\mathbb{P}(M \geq \eta) = \mathbb{P}(M \leq -\eta) \geq \mathbb{P}(M = \eta) = \frac{v}{2n\eta^2}\left(1 - \frac{v}{n^2\eta^2}\right)^{n-1}.$$

7.6. *Proof of Proposition 6.3 (page 1172)*

Let us consider for Y the following distribution, with support $\{-n\eta, -\xi, \xi, n\eta\}$, where ξ and η are two positive real parameters, to be adjusted to obtain the desired variance and kurtosis.

$$\mathbb{P}(Y=-n\eta)=\mathbb{P}(Y=n\eta)=q,$$
$$\mathbb{P}(Y=-\xi)=\mathbb{P}(Y=\xi)=\frac{1}{2}-q.$$

In this case

$$m=0,$$
$$v=(1-2q)\xi^2+2qn^2\eta^2,$$
$$\kappa v^2=(1-2q)\xi^4+2qn^4\eta^4.$$

Thus

$$\kappa=f_q(n\eta/\xi),$$

where

$$f_q(x)=\frac{1-2q+2qx^4}{(1-2q+2qx^2)^2},\quad x\geq 1.$$

It is easy to see that f_q is an increasing one to one mapping of $(1,+\infty($ into itself. We obtain

$$\xi=\sqrt{\frac{v}{1-2q+2q[f_q^{-1}(\kappa)]^2}}\leq\sqrt{v}\quad\text{and}\quad\eta=\frac{\xi f_q^{-1}(\kappa)}{n}.$$

Consequently

$$\left(\frac{\kappa}{2q}\right)^{1/4}\frac{\sqrt{v}}{n}\geq\eta\geq\left(\frac{\kappa-1+2q}{2q}\right)^{1/4}\frac{\sqrt{v}}{n}.$$

On the other hand,

$$\begin{aligned}\mathbb{P}(M\geq\eta-\gamma)&\geq\sum_{i=1}^n\mathbb{P}\left(Y_i=n\eta,\frac{1}{n}\sum_{j,j\neq i}Y_j\geq-\gamma,Y_j\in\{-\xi,\xi\},j\neq i\right)\\&=n\mathbb{P}(Y_1=n\eta)(1-2q)^{n-1}\\&\quad\times\left[1-\mathbb{P}\left(\frac{1}{n}\sum_{i=2}^nY_i\geq\gamma\,\middle|\,Y_j\in\{-\xi,\xi\},2\leq j\leq n\right)\right].\end{aligned}$$

Let us remark that

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\sum_{i=2}^nY_i\geq\gamma\,\middle|\,Y_j\in\{-\xi,\xi\},2\leq j\leq n\right)\\&\leq\inf_{\lambda\geq0}\cosh(\lambda\xi/n)^{n-1}\exp(-\lambda\gamma)\leq\inf_{\lambda\geq0}\exp\left(\frac{\lambda^2\xi^2}{2n}-\lambda\gamma\right)\\&=\exp\left(-\frac{n\gamma^2}{2\xi^2}\right)\leq\exp\left(-\frac{n\gamma^2}{2v}\right).\end{aligned}$$

Also, by symmetry

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=2}^{n} Y_i \geq \gamma \,\middle|\, Y_j \in \{-\xi,\xi\}, 2 \leq j \leq n\right) \leq \frac{1}{2}.$$

Thus, as $m = 0$,

$$\mathbb{P}\big(|M-m| \geq \eta - \gamma\big) \geq 2nq(1-2q)^{n-1} \max\left\{\frac{1}{2}, 1 - \exp\left(-\frac{n\gamma^2}{2v}\right)\right\}.$$

Let us put

$$\chi = \min\left\{\frac{1}{2}, \exp\left(-\frac{n\gamma^2}{2v}\right)\right\}$$

and let us assume that $\varepsilon \leq \frac{1}{16}$. Let us put

$$q = \frac{\varepsilon}{n(1-\chi)}\left(1 - \frac{4\varepsilon}{n(1-\chi)}\right)^{-(n-1)} \leq \frac{2\varepsilon}{n(1-\chi)},$$

the last inequality being a consequence of the convexity of $x \mapsto (1-x)^{n-1} \geq 1-(n-1)x \geq 1-nx, 0 \leq x \leq 1$. Let us remark then that

$$\mathbb{P}\big(|M-m| \geq \eta - \gamma\big) \geq \frac{2\varepsilon(1-2q)^{n-1}}{(1-4\varepsilon/(n(1-\chi)))^{n-1}} \geq 2\varepsilon.$$

Thus, with probability at least 2ε,

$$\begin{aligned}|M-m| &\geq \left[\frac{(\kappa - 1 + 2\varepsilon/n)(1-\chi)}{2n\varepsilon}\right]^{1/4}\left(1 - \frac{4\varepsilon}{n(1-\chi)}\right)^{(n-1)/4}\sqrt{\frac{v}{n}} - \sqrt{\frac{2\log(\chi^{-1})v}{n}} \\ &\geq \sup_{0<\chi\leq 1/2}\left[\frac{(\kappa - 1 + 2\varepsilon/n)(1-\chi-4\varepsilon)}{2n\varepsilon}\right]^{1/4}\sqrt{\frac{v}{n}} - \sqrt{\frac{2\log(\chi^{-1})\mathbb{1}(\chi < 1/2)v}{n}}.\end{aligned}$$

In order to obtain Proposition 6.3 (page 1172), it is enough to restrict optimization with respect to χ to the two values

$$\chi = \frac{(n\varepsilon)^{1/4}}{2} \quad \text{and} \quad \chi = \frac{1}{2}.$$

8. Generalization to non-identically distributed samples

The assumption that the sample is identically distributed can be dropped in Proposition 2.4 (page 1153). Indeed, assuming only that the random variables $(Y_i)_{i=1}^n$ are independent, meaning that their joint distribution is of the product form $\bigotimes_{i=1}^n \mathbb{P}_i$, we can still write, for $W_i = \pm\alpha(Y_i - \theta)$,

$$\begin{aligned}&\mathbb{E}\left\{\exp\left[\sum_{i=1}^{n}\log\left(1 + W_i + \frac{W_i^2}{2}\right)\right]\right\} \\ &\quad= \exp\left\{\sum_{i=1}^{n}\log\left[1 + \mathbb{E}(W_i) + \frac{\mathbb{E}(W_i^2)}{2}\right]\right\} \\ &\quad\leq \exp\left\{n\log\left[1 + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(W_i) + \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\big(W_i^2\big)\right]\right\}.\end{aligned}$$

Starting from these exponential inequalities, we can reach the same conclusions as in Proposition 2.4 (page 1153), as long as we set

$$m = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) \quad \text{and} \quad v = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[(Y_i - m)^2\big].$$

We see that the mean marginal sample distribution $\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_i$ is playing here the same role as the marginal sample distribution in the i.i.d. case.

9. Experiments

9.1. *Mean estimators*

Theoretical bounds can explore high confidence levels better than experiments, and have also the advantage to hold true in the worst case. They have led us to introduce new M-estimators, and in particular the one described in Proposition 4.5 (page 1165). Nonetheless, they may be expected to be rather pessimistic and are clearly insufficient to explore the moderate deviations of the proposed estimators. In particular it would be interesting to know whether the improvement in high confidence deviations has been obtained as a trade-off between large and moderate deviations (by which we mean a trade-off between the left part and the tail of the quantile function of $|\widehat{\theta} - m|$).

This is a good reason to launch into some experiments. We are going to test sample distributions of the form

$$\sum_{i=1}^{d} p_i \mathcal{N}\big(m_i, \sigma_i^2\big),$$

where $d \in \{1, 2, 3\}$, $(p_i)_{i=1,\dots,d}$ is a vector of probabilities and $\mathcal{N}(m, \sigma^2)$ is as usual the Gaussian measure with mean m and variance σ^2. To visualize results, we have chosen to plot the quantile function of the deviations from the true parameter, that is the quantile function of $|\widehat{\theta} - m|$, where $\widehat{\theta}$ is one of the mean estimators studied in this paper.

Let us start with an asymmetric distribution with a so to speak intermittent high variance component. Let us take accordingly

$$\begin{aligned} p_1 &= 0.7, & m_1 &= 2, & \sigma_1 &= 1, \\ p_2 &= 0.2, & m_2 &= -2, & \sigma_2 &= 1, \\ p_3 &= 0.1, & m_3 &= 0, & \sigma_3 &= 30. \end{aligned}$$

In this case, $m = 1$, $\kappa = 27.86$ and $v = 93.5$, so that, when the sample size n is in the range $100 \le n \le 1000$, the variance estimates we are proposing in this paper are not proved to be valid. For this reason we will challenge the empirical mean with the two following estimates: the estimate $\widehat{\theta}_\alpha$ of Proposition 2.4 (page 1153) (where α is chosen using the true value of v) and a naive plug-in estimate, $\widehat{\theta}_{\widehat{\alpha}}$, where $\widehat{\alpha}$ is set as was α, replacing the true variance v with its unbiased estimate $\widehat{V}$ given by equation (4.1), page 1158. The parameter ε is set to $\varepsilon = 0.05$ for both estimators, targeting the probability level $1 - 2\varepsilon = 0.9$.

We will plot also the sample median estimator, in this case where the distribution median is different from its mean, to show that robust location estimators for symmetric distributions do not apply here.

When the sample size is $n = 100$, we obtain the following results (computed from 1000 experiments); see Fig. 15. In this first example, the new M-estimators have uniformly better quantiles than the empirical mean, at any probability level. Moreover the variance can be harmlessly estimated from the data when it is unknown. Thus, in this case, the empirical mean is outperformed from any conceivable point of view.

Let us now increase the sample size to $n = 1000$ (Fig. 16). As should be expected, the values of the three estimators get close for this larger sample size (whereas it becomes more obvious that the empirical median is estimating something different from the mean).

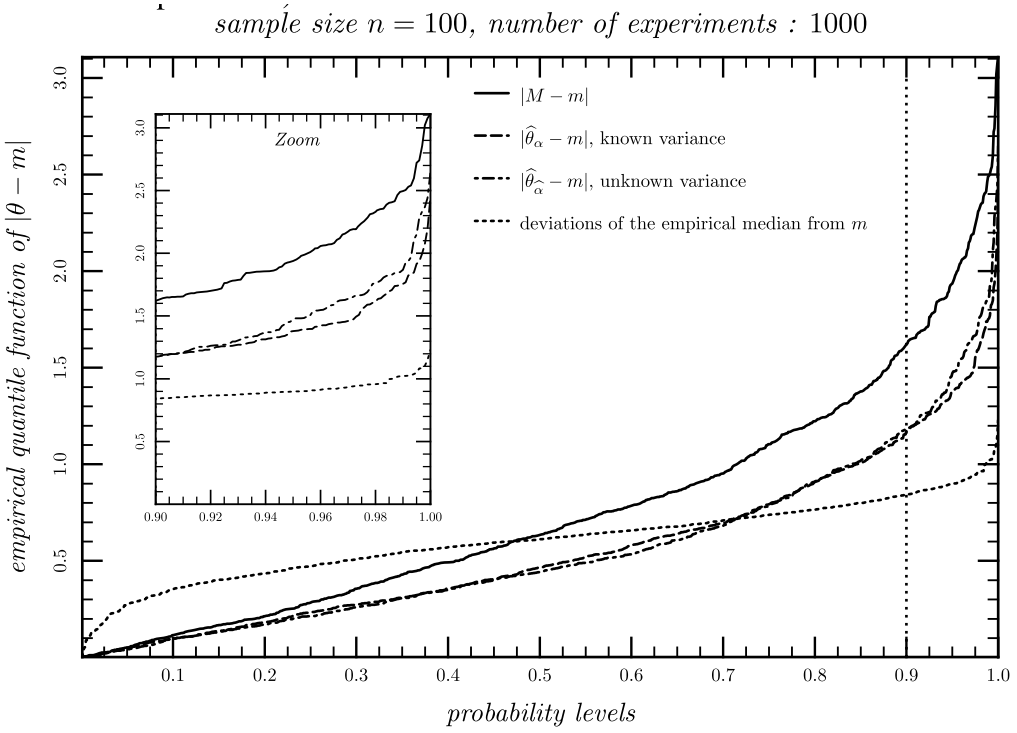

Fig. 15. Some experimental comparison of mean estimators, for a small sample size with a high kurtosis.

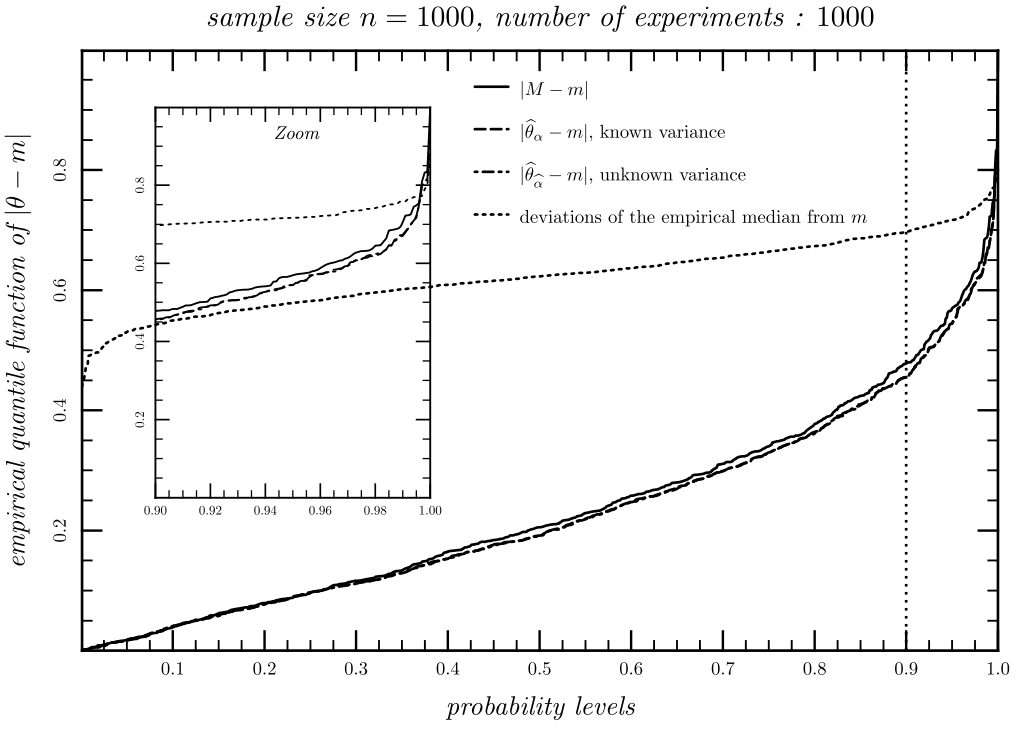

Fig. 16. Experimental study of mean estimators for a larger sample size.

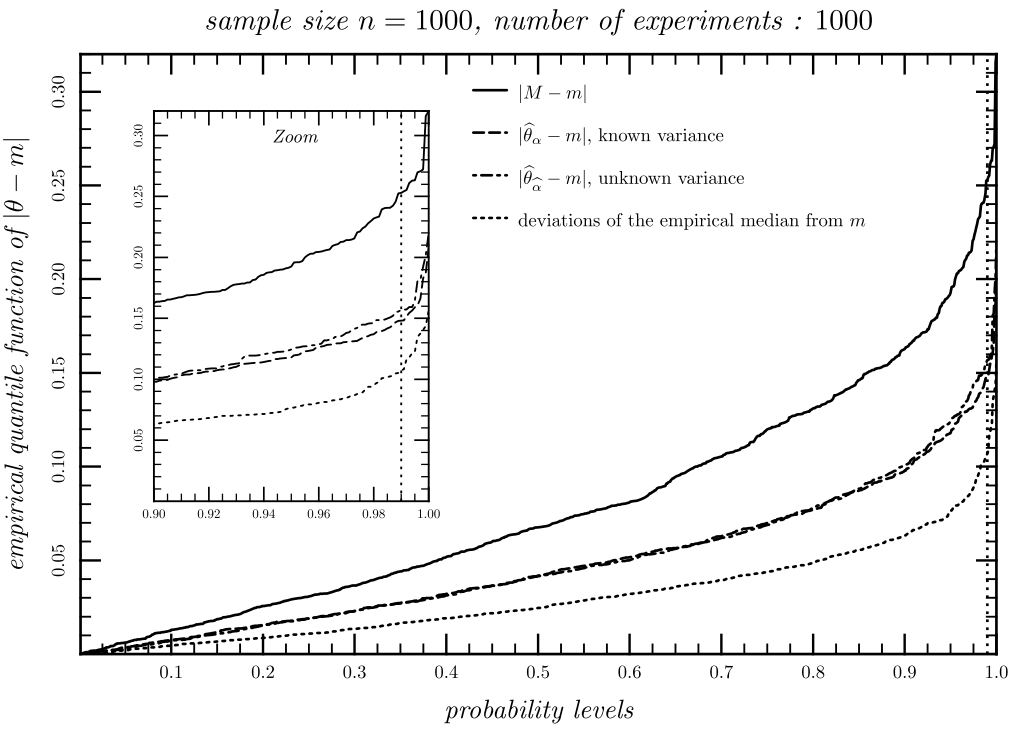

Fig. 17. For each sample size, there are sample distributions for which the empirical mean can be improved.

The empirical mean can still be challenged for this larger sample size (Fig. 17), but for a different sample distribution. To illustrate this, let us consider a situation as simple as the mixture of two centered Gaussian measures. Let $d = 2$ and

$$p_1 = 0.99, \qquad m_1 = 0, \qquad \sigma_1 = 1,$$
$$p_2 = 0.01, \qquad m_2 = 0, \qquad \sigma_2 = 30.$$

Here $\kappa = 243.5$, $v = 9.99$ and $m = 0$. We take $\varepsilon = 0.005$, targeting the probability level $1 - 2\varepsilon = 1 - 10/n = 0.99$.

Let us show some heavily asymmetric situation where the left-hand side of the quantile function of the new estimators does not improve on the empirical mean. In what follows $\kappa = 33.4$, $v = 72.25$, $m = -1.3$, and the mixture parameters are

$$p_1 = 0.94, \qquad m_1 = 0, \qquad \sigma_1 = 1,$$
$$p_2 = 0.01, \qquad m_2 = 20, \qquad s_2 = 20,$$
$$p_3 = 0.05, \qquad m_3 = -30, \qquad s_3 = 20.$$

In Fig. 18 we plot two estimators using the value of the variance, optimized for the confidence level $1 - 2\varepsilon$ with $\varepsilon = 0.05$ and $\varepsilon = 0.0005$ respectively (the estimators with unknown variance show the same behaviour).

Here, choosing a moderate value of the estimator parameter ε is required to improve uniformly on the empirical mean performance, whereas higher values of ε produce a trade-off between low and high probability levels. Whether this remains true in general would require to be confirmed by more extensive experiments.

Let us end this section with a Gaussian sample.

When the sample is Gaussian, as could be expected, our new M-estimators coincide with the empirical mean (Fig. 19). What we obtained for a sample size $n = 1000$ could also be observed for other sample sizes. The deviations of the empirical median are higher in the Gaussian case, as proved in Proposition 6.1 (page 1170) (stating that the deviations of the empirical mean of a Gaussian sample are optimal).

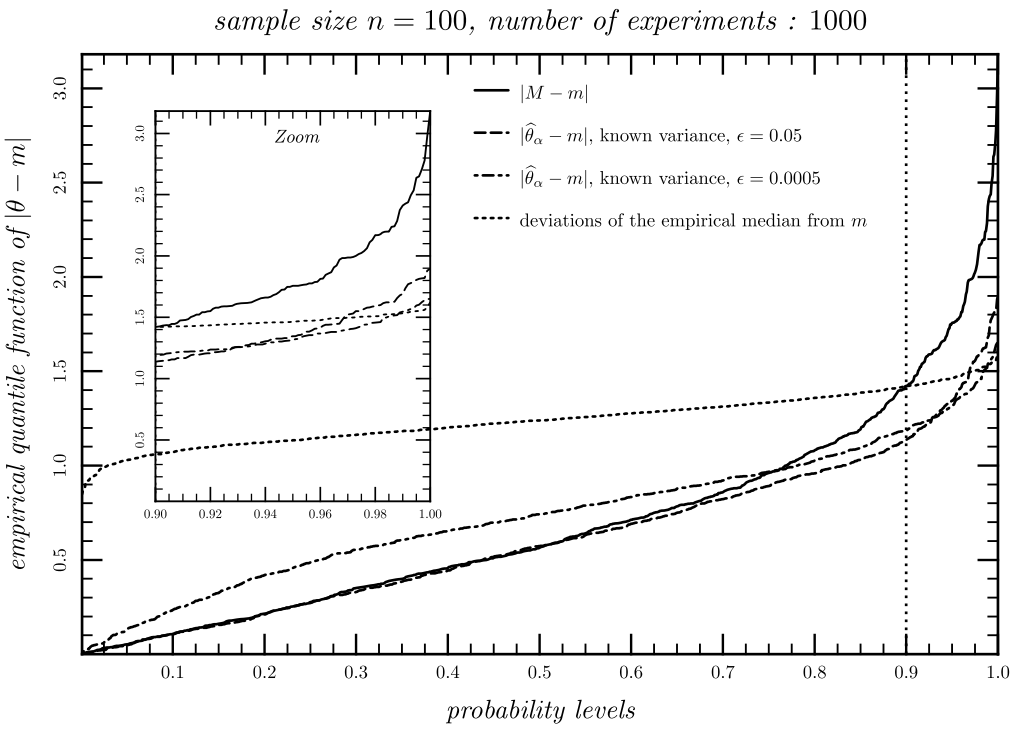

Fig. 18. The trade-off between high and low confidence levels appears when one is targeting very high confidence levels.

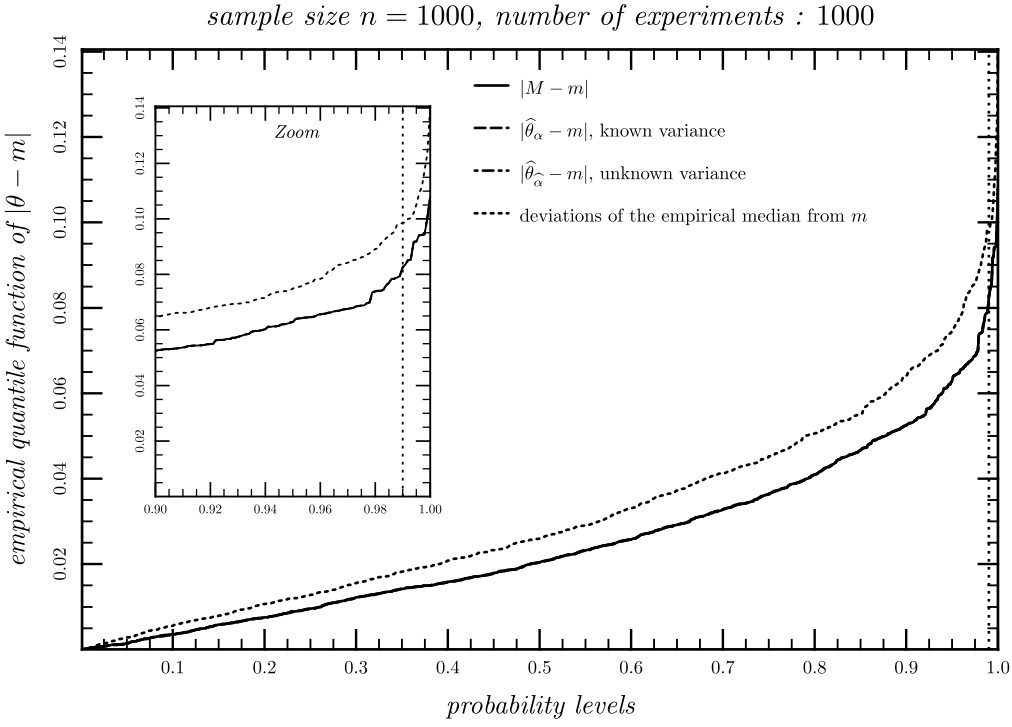

Fig. 19. When the sample distribution is Gaussian, our modified estimator is the same as the empirical mean, which is known to have optimal deviations.

9.2. *Variance estimators*

Let us now test some variance estimates. This is an example where the usual unbiased estimate $\widehat{V}$ defined by equation (4.1), page 1158, shows its weakness. To demonstrate things on simple sample distributions, we choose again a mixture of two Gaussian measures

$$p_1 = 0.995, \qquad m_1 = 0, \qquad \sigma_1 = 1,$$
$$p_2 = 0.005, \qquad m_2 = 1, \qquad \sigma_2 = 5.$$

Here $\kappa = 10.357$, $v = 1.125$ and $m = 0.005$. To be in a situation where the variance estimates of Proposition 4.1 (page 1160) work at high confidence levels, we choose a sample size $n = 2000$, and use in the estimator the parameters $\kappa_{\max} = 6 * n/1000 = 12$, $p = 2$ and $\varepsilon = 0.0025$ (targeting the probability level $1 - 2\varepsilon = 1 - 10/n = 0.0995$); see Fig. 20. So, for the variance as well as for the mean, there are simple situations in which our new estimators perform better in practice than the more traditional ones.

9.3. *Computation details*

To compute the estimators in these experiments, we used the two following iterative schemes (performing two iterations was enough for all the examples shown).

$$\theta_0 = M, \qquad \theta_{k+1} = r(\theta_k) + \theta_k,$$
$$\beta_0 = \frac{\delta - y}{\widehat{V}}, \qquad \beta_{k+1} = \frac{\delta - y}{Q(\beta_k) + \delta}\beta_k.$$

They are based on two principles: they have the desired fixed point and their right-hand side would be independent respectively of θ_k and β_k if ψ was replaced with the identity. The fact that ψ is close to the identity explains why the convergence is fast and only a few iterations are required.

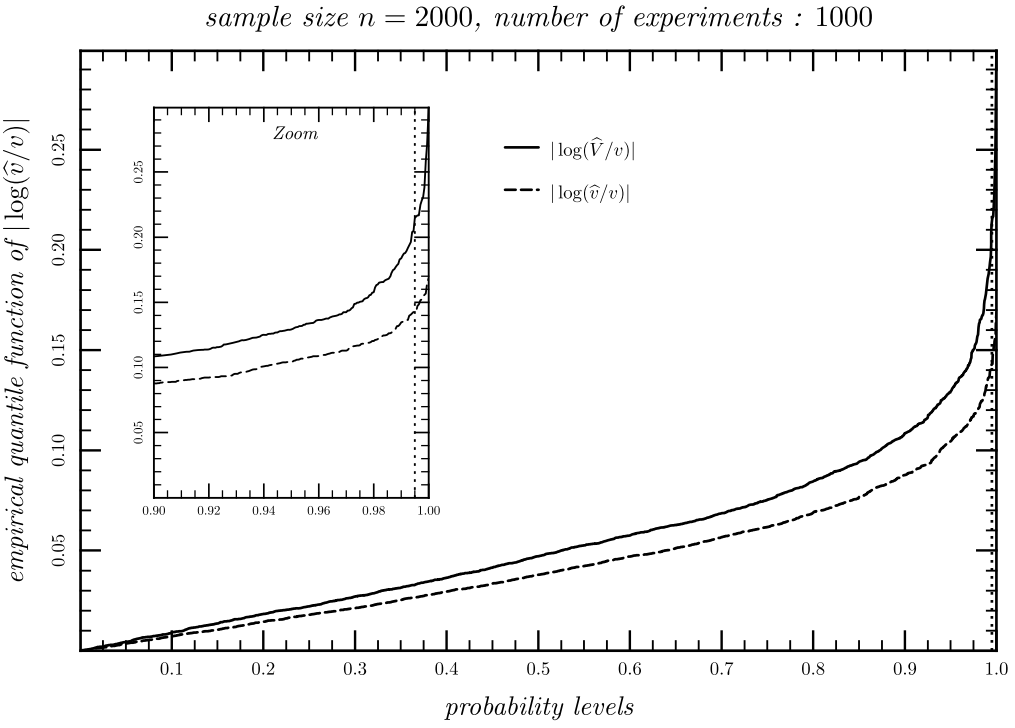

Fig. 20. Experimental study of an improved variance estimate.

These numerical schemes involve only a reasonable amount of computations, opening the possibility to use the new estimators in improved filtering algorithms in signal and image processing (a subject for future research that will not be pushed further in this paper).

10. Conclusion

Theoretical results show that, for some sample distributions, the deviations of the empirical mean at confidence levels higher than 90% are larger than the deviations of some well chosen M-estimator. Moreover, in our experiments, based on non-Gaussian sample distributions, the deviation quantile function of this M-estimator is uniformly below the quantile function of the empirical mean. The improvement of the confidence interval at level 90% can be more than 25%.

Using Lepski's adapting approach offers a response with proved properties to the case when the variance is unknown. For sample sizes starting from 1000, an alternative is to use an M-estimator of the variance depending on some assumption on the value of the kurtosis. However, it seems that the variance can in practice be estimated by the usual unbiased estimator $\widehat{V}$, defined by equation (4.1), page 1158, and plugged in the estimator of Proposition 2.4 (page 1153), although there is no mathematical warrant for this simplified scheme.

References

[1] P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Math. Methods Statist.* **17** (2008) 279–304. MR2483458

[2] J.-Y. Audibert. A better variance control for PAC-Bayesian classification. Preprint n.905bis, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004. Available at http://www.proba.jussieu.fr/mathdoc/textes/PMA-905Bis.pdf.

[3] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.* **39** (2011) 2766–2794. MR2906886

[4] J.-Y. Audibert and O. Catoni. Robust linear regression through PAC-Bayesian truncation. Unpublished manuscript, 2010. Available at http://hal.inria.fr/hal-00522536.

[5] R. Beran. An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** (1978) 292–313. MR0518885

[6] P. J. Bickel. On adaptive estimation. *Ann. Statist.* **10** (1982) 647–671. MR0663424

[7] O. Catoni. *Statistical Learning Theory and Stochastic Optimization: École d'Été de Probabilités de Saint-Flour XXXI – 2001*. *Lecture Notes in Math.* **1851**. Springer, Berlin, 2004. MR2163920

[8] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. *IMS Lecture Notes Monogr. Ser.* **56**. Institute of Mathematical Statistics, Beachwood, OH, 2007. MR2483528

[9] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.* **35** (1964) 73–101. MR0161415

[10] P. J. Huber. *Robust Statistics*. *Wiley Series in Probability and Mathematical Statistics.* Wiley-Interscience, New York, 1981. MR0606374

[11] O. Lepski. Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** (1991) 682–697. MR1147167

[12] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*. Morgan Kaufmann, New York, 1999. MR1811612

[13] D. A. McAllester. Some PAC-Bayesian theorems. *Mach. Learn.* **37** (1999) 355–363.

[14] D. A. McAllester. PAC-Bayesian stochastic model selection. *Mach. Learn.* **51** (2003) 5–21.

[15] C. J. Stone. Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** (1975) 267–284. MR0362669