Do we need to estimate the variance in robust mean estimation?

PDF (/pdf?id=CIv8NfvxsX)

Qiang Sun (/profile?id=~Qiang_Sun2)

24 Aug 2023 (modified: 06 Jan 2024) Decision pending for TMLR Everyone Revisions (/revisions?id=CIv8NfvxsX) BibTeX

Edit ▾

Abstract: In this paper, we propose self-tuned robust estimators for estimating the mean of heavy-tailed distributions, which refer to distributions with only finite variances. Our approach introduces a new loss function that considers both the mean parameter and a robustification parameter. By jointly optimizing the empirical loss function with respect to both parameters, the robustification parameter estimator can automatically adapt to the unknown data variance, and thus the self-tuned mean estimator can achieve optimal finite-sample performance. Our method outperforms previous approaches in terms of both computational and asymptotic efficiency. Specifically, it does not require cross-validation or Lepski's method to tune the robustification parameter, and the variance of our estimator achieves the Cram\'er-Rao lower bound. Project source code is available at https://github.com/statsle/automean (https://github.com/statsle/automean).

Submission Length: Long submission (more than 12 pages of main content)

Previous TMLR Submission Url: /forum?id=NGa8UryQ2F (/forum?id=NGa8UryQ2F)

Changes Since Last Submission:

This is the camera ready version. Specifically,

1. We proof-read the manuscript and the appendix to minimize typos and minor errors.
2. We removed all colored texts.
3. Following the suggestion of one reviewer, we made the constants in Theorem 3.4 and Theorem 3.5 universal.
4. We also made some slight modification to the figures.
5. We made source code available at Github: https://github.com/statsle/automean (https://github.com/statsle/automean).
6. We added a new figure - Figure 1 - in the main text, and added the following discussion: In summary, our self-tuned estimator can achieve optimal performance in both finite-sample and large-sample regimes. We point out that the large-sample regime is used to approximate the regime when the sample size is relatively large instead of describing the case of $n = \infty$. We will refer to this ability as adaptivity to both finite-sample and large-sample regimes, or simply adaptivity. The

MoM estimator does not naturally possess this adaptivity due to its discontinuous nature. Figure 1 provides a comparison between our self-tuned estimator and the MoM estimator in terms of adaptivity.

Competing Interests: 👁 No

Human Subjects Reporting: 👁 No

Code: https://github.com/statsle/automean (https://github.com/statsle/automean)

Assigned Action Editor: Pierre Alquier (/profile?id=~Pierre_Alquier1)

Submission Number: 1503

Search keywords...

Sort: Newest First

👁 Everyone | Authors | Editors In Chief | Action Editors | ✖

21 / 21 replies shown

Add: **Withdrawal** **Official Comment**

Decision by Action Editor 9PEu

Decision ✎ Action Editor 9PEu 📅 24 Oct 2023, 22:01 (modified: 25 Oct 2023, 14:32) 👁 Everyone

Revisions (/revisions?id=uTHXClG0GX)

Claims And Evidence:

This paper is in the line of Catoni (2012) and studies estimators of the mean of a random variable. Catoni's original estimator depends on a tuning parameter which optimal value depends on the unknown variance of the distribution. Many adaptive procedure were proposed, based on Lepski's method among others. Other popular methods such as MOM (Median-of-Means) also depend strongly on tuning parameters. In this paper, the minimization of a new loss function is proposed. This loss function depends both on a mean parameter, and a robustification parameter. By minimizing on both parameters simultaneously, we don't need to use selection/adaptive methods for the variance. A striking result is that the estimator reaches the Cramer-Rao bound, something whichis not known for popular methods such as MOM. The estimator also exhibit strong performances in the numerical experiments.

Audience:

Statisticians / ML theorists.

Robustness is obviously an important topic in machine learning. This paper is more on the statistical / theoretical side, as it focuses on the estimation of the mean of a univariate random variable, so it should be of interest mostly to statisticians and theoreticians of ML. On the other hand, I will point out that MOM methods were also developped for univariate variables, and were extended to multivariate models or model selection. Thus, such foundational papers on robust statistics might definitely be impactful for practitioners in robust ML in the future.

Recommendation: Accept as is

Comment:

All reviewers agreed that the paper "is interesting and appears correct" [5VdC], see also [p9aB]: the topic is of "wide interest", [xmsE]: "This paper provides rigorous theoretical justification on the performance of the proposed method (Theorem 4.2) and its advantage over MoM (Theorem 4.1)". They pointed out that "the paper is well written" [5VdC] and the results are "clear and easy to understand" [xmsE]. On the other hand, some of them were concerned that the results might be too narrow, especially the focus on univariate variables [xmsE]. Also, [5VdC] whether it is possible to calibrate MOM to reach the Cramer-Rao bound.

The discussion with the authors was constructive. A few typos / minor technical details pointed out by the three reviewers (especially [5VdC,p9aB]) were fixed / clarified by the author -- I think this really improved the paper. The authors provided a reason why it might be difficult to reach Cramer-Rao with MOM, and also defended the importance of the results in the unidimensional case. I agree with them: see my comments in the "Audience" box above. Althouth with various level of enthusiasm, the three reviewers recommended to accept the paper and I (enthusiastically!) support their recommendation.

I thank once more the reviewers who were very responsive, this led to a constructive discussions with the author.

Add: **Official Comment**

Review of Paper1503 by Reviewer p9aB

Review Reviewer p9aB 25 Sept 2023, 01:37 (modified: 25 Sept 2023, 01:37) Everyone

Revisions (/revisions?id=9YmGOLF3OD)

Summary Of Contributions:

This work studies the problem of how to estimate the mean of a distribution with finite variance given i.i.d. samples and proposes a self-tuned estimator which does not require cross-validation or Lepski's method to tune hyper-parameter. The estimator is based on jointly minimizing a newly defined penalized pseudo-Huber loss function over both the estimation variable μ and robustness variable ν. For finite-sample theory, an estimation error bound in the order of $O(\sqrt{\log(n)/n})$ is established, where n is the number of samples.To compare with the existing estimators median-of-means (MoM) and trimmed mean estimator, the asymptotic efficiency of the proposed estimator is studied. In numerical experiments, the proposed method achieves lower estimation errors on datasets generated from skewed generalized t distributions, compared with sample mean, MoM, trimmed mean, cross validation, and Lepski's method, and it is also more computationally efficient than the last two methods.

Strengths And Weaknesses:

Strengths

(1) The problem of mean estimation of skewed or heavy tailed distributions is fundamental and of wide interest. Thus, an efficient method that is not overly complicated could have a broad impact.

(2) The main idea of the proposed self-tuned estimator, jointly minimizing over estimated mean μ and robustness parameter ν, is clearly explained. Theorem 2.3 further justifies minimization over ν by proving that, given ground truth $\mu_\star$, the optimal $\nu_\star$ can converge to the distribution variance σ as n tends to infinity.

(3) Conclusions on finite-sample estimation error bound and asymptotic error bound provide theoretical guarantees on the performance of the estimators. These theoretical results also justify the advantage of the proposed estimator over existing methods.

Weaknesses/questions

(1) The motivation of defining the penalized pseudo-Huber function is to avoid trivial solutions 0 and ∞. Then why is the constraint $\nu_0 \leq \nu \leq V_0$ still needed to guard $\hat{\nu}$ from 0 and ∞ in (3.1)? Since problem (3.1) is the actual optimization problem to be solved, why not just use the pseudo-Huber function? How do we choose ν_0 and V_0 in practice? Is cross-validation still needed?

(2) The theorems can be stated more clearly.

- In Theorem 3.1, both z and r depend on δ, but after the definitions of z and r it says that for any δ the error bound holds with probably $1-\delta$. Are they the same δ? If so, it'd be better to put the phrase ``for any δ'' in the front, before defining z and r. The same issue is also in Lemma 3.2.
- In Theorem 3.1, the error bound actually equals to the radius r. It'd be more clear to use the same notation and state $|\hat{\mu} - \mu^\star| < r$. Also, is there any intuitive explanation on why the error bound equals to the radius in the assumption?
- In Lemma 3.2 and Corollary 3.3, one assumption is $n \geq C \max(z^2(\sigma^2 + r^2)/v_0^2, \log(1/\delta))$, but r also depends on n, so what is the requirement on n for this assumption to hold? It would be better to plug r into this inequality to draw a condition on n.

(3) In Theorem 3.4 and Theorem 3.5, one assumption is $c_0\sigma_{v_0^2 n/z^2} \leq C_0\sigma$. What is the requirement on n for this assumption to hold? It'd be better to discuss this issue at least for some commonly encountered distributions. The key question here is whether or not this assumption imposes a more strict condition on n.

(4) In the numerical section, what is the method/solver actually used to solve problem (3.5)? How does this method/solver scale as the number of samples n increase? This would determine the runtime efficiency of the proposed estimator.

(5) In the numerical section, is there a reason not to compare all methods in the same plot? For example, can Figure 1 and Figure 3 be merged together? It'd be better to also report the runtimes of sample mean, MoM, and trimmed mean.
Requested Changes:
The points in the Weaknesses/questions are my suggested adjustments. In my view, (1) (3) (4) are critical, while the others would strengthen the work.

Besides, there are some minor issues.

- Should the $\hat{\mu}(\tau)$ in (2.2) be $\tilde{\mu}(\tau)$?
- Is the $\hat{\mu}(\tau)$ in Theorem 2.1 the solution of the pseudo-Huber loss or the penalized pseudo-Huber loss?
- In the last paragraph of section 2, what is the loss function ℓ? Should it be ρ instead?

Broader Impact Concerns:
I have no concern on the ethical implications of the work.

Claims And Evidence: Yes
Audience: Yes

Add: Official Comment

Response to reviewer p9aB, part 1

Edit

Official Comment Authors (Qiang Sun (/profile?id=~Qiang_Sun2))
02 Oct 2023, 23:44 (modified: 25 Oct 2023, 15:32) Everyone Revisions (/revisions?id=elb3ok1Z3u)

Comment:
We would like to thank this reviewer for many helpful comments which have helped improved the presentation of the paper. We address the comments in a one-to-one manner below.

Major comments

> 1. Why is the constraint $v_0 \leq v \leq V_0$ still needed? Since problem (3.1) is the actual problem to be solved, why not just use pseudo-Huber function? How do we choose v_0 and V_0 in practice? Is cross-validation still needed?

We first apologize for the typo in Eq. (3.1). We suspect that most of these questions are due to this typo. As pointed out by reviewer 5Vd3, the $\hat{v}$ in the constraint in Eq. (3.1) should be v. Thus it is NOT equivalent to the pseudo-Huber loss. The v_0 and V_0 in practice are taken as small and large constants. In our experiments, we take v_0 and V_0 as 10^{-4} and 10^4 respectively. The pseudo-Huber loss, when treated as a loss function for v, has trivial minimizers $v = 0$ and $v = \infty$. The penalized pseudo-Huber avoids these two trivial minimizers, aka when starting an initialization, an algorithm such as gradient descent will not converge to these trivial minimizers. Yet, we still need the constraint that $v_0 \leq v \leq V_0$. The reason is that, to run an algorithm such as the alternative gradient descent (see Appendix C), we need to initialize v at some initialization v_{init} bounded away from 0 and ∞, because $v = 0$ and ∞ correspond to a nonsmooth loss function (in μ) and a trivial loss function respectively. And smoothness is needed for establishing the strong convexity of the loss function. Technically speaking, we need $v_0 \leq v \leq V_0$ to bound the tail probability of $\mathcal{E}_2$; see Lemma G.1 and proof of Theorem 3.4.

> 2. (a). It is better to put the phrase “for any δ” in the front before defining z and r in Theorem 3.1 and Lemma 3.2. (b). Theorem 3.1. The error bound actually equals to the radius r. Why is the error equal to r? (c). In Lemma 3.2 and Corollary 3.3, the assumption involves r. Can you make this explicit?

We address these questions as follows.

(a) We have fixed this.

(b). We apologize for this confusion. We have revised the statement of theorem to be “Assume Assumption 1 holds with any $r{\ge}r_0 := (\kappa_\ell)^{-1}\Big(\sigma/(\sqrt{2}v)+1\Big)^2\sqrt{\log(2/\delta)/n}$ .... ” Indeed we need Assumption 1 to hold with an r that is larger than the upper bound, so that we can utilize the strong convexity to convert the loss error bound to the parameter estimation error bound; see the proof of Theorem 3.1. Thus r has be larger than that error bound. Taking r to be as small as possible, i.e., $r = r_0$, results in Assumption 1 being at its weakest.

(c). Because the penalized pseudo-Huber loss function transitions from a quadratic to a linear function when $|x|$ is approximately $\sqrt{n}$, thus Assumption 1 holds with any r such that $\sqrt{n} \gtrsim r \ge r_0$. Thus we can take r to be a constant, and this will not make the sample complexity condition worse. For example, to make the assumption independent of r, we can simply take $r = \sigma$ which, by $r \ge \mathit{error\ bound}$ implies that the following sample complexity condition

$$r{\ge}(\kappa_\ell)^{-1}\Big(\sigma/(\sqrt{2}v)+1\Big)^2\sqrt{\log(2/\delta)/n},$$

which is implied by

$$n \ge \kappa_\ell^{-2}(1/v_0+1/\sigma)^2\log(2/\delta).$$

This, together with $n \ge C\max\{2z^2\sigma^2/v_0^2,\ \log(1/\delta)\}$, gives the sample complexity condition. Directly plugging $r^2 = r_0^2$ into the sample complexity leads to a quadratic inequality in n and solving it gives complicated and hard-to-interpret sample complexity condition. We prefer the current condition because it is intepretable. Specifically, the first sample complexity condition that $n \ge Cz^2(\sigma^2+r^2)/v_0^2$ comes from requirement that $\tau_{v_0}^2 := v_0^2n/z^2 \ge C(\sigma^2+r^2)$ in the proof of Lemma 3.2. Recall that the robustification parameter $\tau_{v_0}^2 := v_0^2n/z^2$ determines the size of the quadratic region. Thus, intuitively, this requirement is minimal in the sense that Assumption can only hold when $\tau_{v_0}^2$ is larger than r^2 plus the noise variance σ^2 (due to stochasticity).

> 3. In Theorem 3.4 and Theorem 3.5. One assumption is $c_0\sigma_{v_0^2n/z^2} \le C_0\sigma < V_0$.

Because $\sigma_{v_0^2n/z^2} = \sqrt{\mathbb{E}[\varepsilon^2 1(\varepsilon^2 \le v_0^2n/z^2)]}$ is the truncated standard deviation, we have $\sigma_{v_0^2n/z^2} \le \sigma$, and thus $c_0\sigma_{v_0^2n/z^2} \le C_0\sigma < V_0$ automatically holds with $c_0 \le C_0$. Following the suggestions of other reviewers, we have rewritten the statements of the results. Please see the revised paper for details.

Add: **Official Comment**

Response part 2

Edit

Official Comment Authors (Qiang Sun (/profile?id=~Qiang_Sun2))
02 Oct 2023, 23:45 (modified: 03 Oct 2023, 00:32) Everyone Revisions (/revisions?id=Cg7nMkJBZr)

Comment:

> 4. In the numerical section, what is the method/solver actually used to solve problem (3.5)? How does this method/solver scale as the number of samples increase? This would determine the runtime efficiency of the proposed estimator.

We use alternating gradient descent with the Barzilai and Borwein method and backtracking line search (in python); see Appendix C for our meta algorithm. For the same setting as in Table 1, for $n = 100, 1000, 10000, 100000$ (repeated 1000 times), the run time is 1.54, 1.58, 3.02, 25.04 seconds, respectively:

n	100	1000	$10,000$	$100,000$
time	1.54	1.58	3.02	25.04

We will release the code on github publicly.

> 5. In the numerical section, is there a reason not to compare all methods in the same plot? For example, can Figure 1 and Figure 3 be merged together? It'd be better to also report the runtimes of sample mean, MoM, and trimmed mean.

We provide the runtimes of sample mean, MoM, and trimmed mean. Specifically. When combining Figure 1 and Figure 3 into one figure, we found that the figure got every crowded with 4 lines overlapping together and it is very hard to distinguish different lines; see the lefts panels of Figure 1 and Figure 3. Moreover, Figure 3 is all for minimizing the penalized pseudo-Huber loss with different methods for tuning the robustification parameter. Thus we prefer not to combine Figure 1 and Figure 3. Table 1 is mainly for comparing the run time for different tuning methods. We report the runtime of sample mean, MoM, and trimmed mean here and in the last second paragraph on page 10. Specifically, for the same setting as in Table 1, the run time for sample mean, MoM, and trimmed mean is 0.018, 0.111, and 0.057, respectively:

method	**sample mean**	**MoM**	**trimmed mean**
time	0.018	0.111	0.057

Minor comments:

> 1. Should the $\hat{\mu}(\tau)$ in (2.2) be $\tilde{\mu}(\tau)$?

We have fixed this. The earlier $\tilde{\mu}(\tau)$ should be $\hat{\mu}(\tau)$.

> 2. Is the $\hat{\mu}(\tau)$ in Theorem 2.1 the solution of the pseudo-Huber loss or the penalized pseudo-Huber loss?

We have fixed the previous one and thus this one. It should be the solution to the pseudo-Huber loss.

> 3. In the last paragraph of section 2, what is the loss function ℓ? Should it be ρ instead?

Yes, we have fixed this. Thanks for pointing this out.

Add: Official Comment

Replying to Response part 2

Thank you

Official Comment Reviewer p9aB 23 Oct 2023, 05:12 Everyone

Comment:
Thank you for responding to my comments in great details.

Add: Official Comment

Review of Paper1503 by Reviewer 5VdC

Review Reviewer 5VdC 14 Sept 2023, 22:23 (modified: 25 Sept 2023, 01:37) Everyone

Revisions (/revisions?id=6FCXZHeUjw)

Summary Of Contributions:
This paper addresses the question of computationally efficient, heavy-tailed, one-dimensional mean estimation for distributions with finite but unknown covariance. A new estimator is proposed for this task, which adapts to the unknown variance and achieves near optimal finite-sample performance. Their estimator is shown to be asymptotically efficient (achieving the CR lower bound), in comparison to existing approaches for the problem, and can be computed efficiently. Numerical experiments demonstrating the superiority of the proposed approach compared to other methods are also provided.

The idea behind the construction is the following: the Huber loss can be used for this task with known variance (i.e., penalize mean estimate by expected squared loss for small loss values and mean absolute deviation for large loss values). Further, it is possible to smooth/convexify the Huber loss, which yields the pseudo-Huber loss. Both depend on some proxy for the variance, which is unknown. The authors then add in an extra optimization parameter for this variance proxy and augment the pseudo-Huber loss so that the expected risk is minimized when this parameter is a decent proxy of the true variance.

Strengths And Weaknesses:
Strengths:

- The self-tuning estimator and the fact that it overcomes the unknown variance issue is interesting to me, and I wonder if this can be characterized as a special case of a more general approach. The idea is novel (I, at least, have not seen this approach before) and constitutes a good contribution.
- The paper is well written. The technical writing is solid, the exposition/motivation/outlook are to a good level as well. I enjoyed reading it.
- The efficiency result provides a nice theoretical distinction between the proposed estimator and the MoM method. The argument is reminiscent of that for the classical sample mean vs. median estimator for the expectation (it seems appropriate to mention this and provide a brief discussion). This result provides a theoretical explanation of the improved performance observed empirically.
- Speaking of the experiments, the practical performance of the estimator is solid, beating others consistently (though see second weakness, not clear that they couldn’t be improved)

Weaknesses:

-The conditions for the finite-sample guarantees from Theorems 3.4 and 3.5 seem quite subtle and are hard to interpret. Could these conditions be relaxed to more natural and primitive assumptions?

- The MoM estimator is dead simple, more computationally efficient, and solves the formulated problem. The authors admit this but emphasize the asymptotic performance as a key benefit of their approach. However, the guarantees of the self-tuning approach depend on some subtle assumptions, and the finite-sample performance depends on unknown constants. This means that the sample complexity cannot be computed in advance for fixed error. This should be mentioned in the limitation section.
- A glaring limitation of the approach is the fact that it only treats the one-dimensional case. I find it odd that the authors do not even comment on the extension to $d > 1$. For example, they can apply the current method coordinate-wise to obtain an L^2 error of $\sqrt{d}$ times the individual error, perhaps with some $\log d$ factor for a union bound. I encourage the authors to examine this case and, if possible, provide an analysis. The attained risk may not be optimal but still worth mentioning. Either way, the limitations bit in the Conclusion section should discuss the extension to higher dimensions and the roadblocks towards it.
- It is not clear that the MoM and trimmed mean estimators have been optimized for asymptotic efficiency. I wonder if the constant of 8 which appears in the MoM case could be reduced for large n in a way that would make it more efficient. This idea is implemented in the proposed approach and leads to its superiority. It seems appropriate to try check where existing approaches can be adapted or prove a formal limitation result.
- My sense is that this approach is unlikely to scale to high-dimensions (in general, M-estimation like Huber loss usually does not work for high-dimensional robust statistics). Did the authors try looking at the $d > 1$ case. The authors motivate their work by mentioning `high-dimensional data analysis' in the first sentence of the intro. It would be appropriate to the multivariate case or at least adequately discuss it and explain the roadblocks towards the extension. This too should be added to the limitations section.

Overall:

Despite the above weaknesses, the paper is interesting and appears correct. If the authors could get finite-sample performance guarantees with more interpretable assumptions, and/or argue convincingly that MoM or similar approaches could not be easily improved, I think it will qualify for acceptance.

Requested Changes:
Beyond addressing the weaknesses above, here are a few more secondary comments:

- I suggest setting $\alpha = 0.5$ from the get go and perhaps comment on why this is the right choice. The current approach (which leaves it as a free parameter and then arrives at this choice at the end of the paragraph following Theorem 2.3) does not contribute to clarity nor generality.
- In regard to Theorems. 4.2-4.3, do the authors have intuition about what is the distinctive characteristic of the proposed approach that enables its asymptotic efficiency, compared to the MoM estimator? Adding such a discussion would be a welcome addition. Another small point: I suggest being consistent with how $\iota \in (0, 1]$ (vs. $0 < \iota \leq 1$) is written in those theorems.
- v_0 should be fixed at the start of Lemma 3.2. Currently, it appears out of nowhere.
- In Eq. (3.1), $\hat{v}$ should be just v.
- Can the authors provide an interpretation for the variance ration assumption from Theorem 3.4? Perhaps identifying primitive sufficient conditions or providing an example that verifies it would be helpful.

Broader Impact Concerns:
No concerns.

Claims And Evidence: Yes
Audience: Yes

Add: Official Comment

-
=
≡

Response to reviewer 5VdC, part 1

Edit ▾

Official Comment ✎ Authors (Qiang Sun (/profile?id=~Qiang_Sun2))
02 Oct 2023, 23:51 (modified: 25 Oct 2023, 15:34) Everyone Revisions (/revisions?id=bkXXfTMm9G)

Comment:
We would like to thank this reviewer for many helpful comments which have helped improved the presentation of the paper. We address the comments in a one-to-one manner below.

Major comments:

> Major comment 1 and minor comment 5. The conditions for the finite-sample guarantees from Theorems 3.4 and 3.5 seem quite subtle and are hard to interpret. Could these conditions be relaxed to more natural and primitive assumptions?

Major comment 1 and minor comment 5 are similar, and we address them here. Thanks for pointing this out. We have restated the theorem. Specifically, we have removed this variance ratio condition. Instead, the lower bound in Theorem 3.4 becomes

$$\hat{v} \geq c_0 \sigma_{\tau_{v_0}^2 - \epsilon_0} \geq c_0 \sigma_{\tau_{v_0}^2 - 1}.$$

See our revised manuscript for details. Theorem 3.5 has been revised accordingly.

> 2. However, the guarantees of the self-tuning approach depend on some subtle assumptions, and the finite-sample performance depends on unknown constants. This means that the sample complexity cannot be computed in advance for fixed error. This should be mentioned in the limitation section.

Thank you for this suggestion. We have addressed in our limitation section.

> 3&5. The multi-dimensional case $d > 1$.

The 3&5 major comments are on the study of multi-dimensional case $d > 1$, so we address both of them here in this thread. Following the suggestion of this reviewer, we have applied our estimator coordinate-wise and obtained a result for the multivariate case. See the discussion section for details.

> 4. It is not clear that the MoM and trimmed mean estimators have been optimized for asymptotic efficiency. I wonder if the constant of 8 which appears in the MoM case could be reduced for large n in a way that would make it more efficient. This idea is implemented in the proposed approach and leads to its superiority. It seems appropriate to try check where existing approaches can be adapted or prove a formal limitation result.

We extremely thank this reviewer for asking this inspiring question! This question makes us to deep dive the working mechanisms of our estimator and the MoM estimator, as stated below.

We start with explaining intuitively why our self-tuned estimator can achieve (near) optimal performance in both the finite-sample regime and the asymptotic regime. Because the self-tuned estimator in (3.1) is a self-tuned version of the pseudo-Huber estimator in (2.2), we focus on the pseudo-Huber estimator $\hat{\mu}(\tau)$. Theorem 2.1 suggests that taking $\tau = \sigma\sqrt{n/\log(1/\delta)}$ guarantees the sub-Gaussian performance of $\hat{\mu}(\tau)$ for finite samples. Meanwhile, as $n \to \infty$, we have $\tau = \sigma\sqrt{n/\log(1/\delta)} \to \infty$. Thus the pseudo-Huber loss approaches to the least square loss which corresponds to the negative log maximum likelihood of Gaussian distributions, which leads to the asymptotically efficient mean estimator.

For MoM estimators, the situation differs. On one hand, to attain robustness in the finite-sample regime, the number of blocks k should be greater than or equal to $\lceil 8\log(1/\delta) \rceil$, as demonstrated in the proof of Theorem 4.1 by Lugosi and Mendelson (2019). On the other hand, to approach the sample mean estimator and achieve asymptotic efficiency in the large sample limit, the number of blocks should diminish to 1 as the sample size n

grows. Consequently, optimal finite-sample and asymptotic properties represent two contrasting characteristics for MoM estimators. In other words, the MoM estimator can not simultaneously adapt to both regimes. This contrast seems to arise from the discontinuous nature of the MoM estimator which cannot smoothly transition from requiring at least $k = 3$ blocks (for defining the median) to functioning as an empirical mean estimator.

References:

Lugosi and Mendelson (2019), Mean estimation and regression under heavy-tailed distributions — A survey.

Add: **Official Comment**

-
=
≡

Response part 2

Edit ▾ 🗑

Official Comment ✎ Authors (👁 Qiang Sun (/profile?id=~Qiang_Sun2))

02 Oct 2023, 23:52 (modified: 25 Oct 2023, 15:31) 👁 Everyone Revisions (/revisions?id=EdlE0rh7Im)

Comment:
Minor comments:

> 1. I suggest setting $\alpha = 0.5$ from the get go and perhaps comment on why this is the right choice.

Thanks for pointing this out. We thought about this. But with that change, it is hard to discuss the optimality of $a = 1/2$ and it is even harder to connect to Huber's concomitant estimator discussed in the end of Section 2. After some reflection, we prefer to keep as it is in Section 2. For all discussions after Section 2, we set $a = 1/2$.

> 2. In regard to Theorems. 4.2-4.3, do the authors have intuition about what is the distinctive characteristic of the proposed approach that enables its asymptotic efficiency, compared to the MoM estimator? Adding such a discussion would be a welcome addition. Another small point: I suggest being consistent with how $\iota \in (0, 1]$ (vs. $0 < \iota \leq 1$) is written in those theorems.

For the intuition, please see our comment to major comment 4. We have also changed all $\iota \in (0, 1]$ to $0 < \iota \leq 1$ to make it consistent.

> 3.v_0 should be fixed at the start of Lemma 3.2.

Thank you, we have fixed this.

> 4. In Eq. (3.1), $\hat{v}$ should be just v.

Thank you, we have fixed this.

> minor comment 5

This is addressed in response to major comment 1.

Add: Official Comment

Official Comment by Reviewer 5VdC

Official Comment Reviewer 5VdC 09 Oct 2023, 21:29 (modified: 09 Oct 2023, 21:30) Everyone
Revisions (/revisions?id=TfvJ5vHK46)

Comment:
Thank you for the response to the above points my comments and questions. Below, please find some additional thoughts:

- In the authors' repones to my original "Major comment 1 and minor comment 5", I don't see how this modification addresses the original concern. Perhaps the authors can provide more context?
- I like the explanation for the asymptotic efficiency and believe that including this discussion has improved clarity and interpretability of the results. Thank you for adding that.
- The sample complexity results leave much to be desired. Having bounds that *explicitly* depend on the parameters of the problem would significantly strengthen them. Perhaps the c_0 and C_0 constants can be bounded within certain ranges? Is it clear that such constants exist for any parameter values? I don't view a full account of the above as mandatory for acceptance; in my opinion, the paper has merit and presents innovative ideas even with this limitation. Still, if more can be said, that would be welcome. (unrelated: I also suggest changing the phrasing in Thms. 3.4 and 3.5 from "Let c0 and C0 be some constants, and suppose ..." to "Suppose that c0 and C0 are constants with [or: such that]...".)
- I am pleased with the remaining responses. Thank you.

Add: Official Comment

Further response to reviewer 5VdC

Edit

Official Comment Authors (Qiang Sun (/profile?id=~Qiang_Sun2))
09 Oct 2023, 22:34 (modified: 11 Oct 2023, 13:19) Everyone Revisions (/revisions?id=PkdPHasuae)

Comment:
Thanks for your positive responses and follow-up questions! We address two of your questions below. We start with comment 3.

> 3. The sample complexity results leave much to be desired. Having bounds that *explicitly* depend on the parameters of the problem would significantly strengthen them. Perhaps the c_0 and C_0 constants can be bounded within certain ranges?

This is a good suggestion. c_0 and C_0 here have simple dependencies on problem-dependent quantities. But there are also other constants that depends on other constants and problem-dependent quantities in a complicated way. For readability concerns, we hide all the dependencies. In hindsight, we agree that having bounds that explicitly depend on the parameters might be useful. To balance the readability and explicitness of the results, in the follow-up revision, we plan to provide results with explicit constants in the appendix. We will also revise the phrasing in the results as you suggested.

> 1. In the authors' repones to my original "Major comment 1 and minor comment 5", I don't see how this modification addresses the original concern. Perhaps the authors can provide more context?

We apologize if we misunderstood your comment. In our original submission, our statement of Theorem 3.4 reads as “Assume that n is sufficiently large and there exists an $\epsilon_0 = \tilde{O}(1/n)$ such that $\sigma_{\tau_{v_0}^2/2-\epsilon_0}/\sigma_{\tau_{v_0}^2/2} \geq 1/3$....”

We agree that requiring this variance ratio assumption is pretty strong especially with the constant $1/3$ there. We modified the proof, and removed this assumption. But now the lower bound depends on $\sigma_{\tau_{v_0}^2/2-1}/\sigma_{\tau_{v_0}^2/2}$. We clarify that this is not a very strong assumption as $\tau_{v0}^2 \geq \tau_{v0}^2/2 - 1 \to \infty$, and thus, by the dominated convergence theorem, we have both $\sigma_{\tau_{v_0}^2/2-1}$ and $\sigma_{\tau_{v_0}^2/2}$ approach to σ, and thus their ratio approaches to 1.

We give an example. Suppose the third absolute moment exists. Then

$$\frac{\sigma^2_{\tau_{v_0}^2/2-1}}{\sigma^2_{\tau_{v_0}^2/2}} = 1 - \frac{\sigma^2_{\tau_{v_0}^2/2-1}}{\sigma^2_{\tau_{v_0}^2/2}} \geq 1 - \frac{\mathbb{E}|\epsilon|^3}{\sigma^2_{\tau_{v_0}^2/2}} \times \frac{1}{\sqrt{\tau_{v_0}^2/2-1}} = 1 - O\left(\frac{1}{\sqrt{n}}\right).$$

We will follow up with a revision that has these added.

Add: Official Comment

Replying to Further response to reviewer 5VdC

Official Comment by Reviewer 5VdC

Official Comment Reviewer 5VdC 11 Oct 2023, 09:23 Everyone

Comment:

Thank you. The planned revisions sound good to me.

Add: **Official Comment**

Replying to Official Comment by Reviewer 5VdC

Thank you

Edit

Official Comment Authors (Qiang Sun (/profile?id=~Qiang_Sun2)) 12 Oct 2023, 15:14 Everyone

Comment:
Thank you very much for your review! We believe your comments have helped to improve the paper significantly! If you have any other questions/confusion, please let us know.

Add: **Official Comment**

Review of Paper1503 by Reviewer xmsE

Review Reviewer xmsE 12 Sept 2023, 13:03 (modified: 25 Sept 2023, 01:37) Everyone

Revisions (/revisions?id=ZM7ZLZlrlJ)

Summary Of Contributions:
This paper proposes an algorithm for robust mean estimation. The proposed method can automatically estimate the variance, and theoretical justification on its advantage over median of means (MoM) is provided. Numerical results also demonstrate its advantage.

Strengths And Weaknesses:
Pros:

This paper provides rigorous theoretical justification on the performance of the proposed method (Theorem 4.2) and its advantage over MoM (Theorem 4.1). The mathematical statements are clear and easy to understand.

Cons:

My two major concerns are about the contribution and the topic of this paper:

- This paper studies a robust mean estimation and provides a lot of theoretical justification. However, in my point of view, this paper is more suitable for a statistics journal rather than a ML journal for the lack of real-data analysis. Mean estimation is a very traditional statistical problem.
- The theory main considers a univariate case and does not consider multi-variate case. A theory in a multi-variate scenario may provide more insights on the relationship between the mean estimation task performance and the covariance structure of the distribution, i.e., for the asymptotic normality results in Theorem 4.1 and 4.2, how are the asymptotic distributions related to the covariance structure?

Besides the two main concerns, the following are some minor comments in writing:

- I would suggest mentioning "heavy-tailed distribution" in either the abstract or the title.
- The first paragraph of the paper is a little bit misleading. It mentions about some applications in high-dimensional statistics. However, starting from the second paragraph, it talks about a mean estimation problem for a univariate random variable.
- In the first sentence of the abstract, "...self-tuned robust estimators for estimating the mean of distributions with only finite variances", does "with only finite variances" refer to the robust estimator, or refer to the distribution?
- In "the resulting estimator for the robustification parameter can adapt to the unknown variance automatically", what does the "variance" refer to?

Requested Changes:
Please provide some more results in multi-variate asymptotic distribution.

Broader Impact Concerns:
NA

Claims And Evidence: Yes
Audience: No

Add: Official Comment

Response to reviewer xmsE

Edit

Official Comment Authors (Qiang Sun (/profile?id=~Qiang_Sun2))
03 Oct 2023, 00:09 (modified: 08 Oct 2023, 17:00) Everyone Revisions (/revisions?id=dddJT9rBce)

Comment:
We would like to thank this reviewer for his comments which have helped the authors to dig deeper about the subject. We address the comments in a one-to-one manner below.

Major comments:

> 1. However, in my point of view, this paper is more suitable for a statistics journal rather than a ML journal for the lack of real-data analysis. Mean estimation is a very traditional statistical problem.

As also pointed out by other reviewers, we argue that the mean estimation problem of skewed or heavy tailed distributions is fundamental and of wide interest to statistics as well as machine learning community. Thus, an efficient method that is not overly complicated could have a broad impact. Our method also serves as the foundation for tackling more general problems.

> 2. The theory main considers a univariate case and does not consider multi-variate case. A theory in a multi-variate scenario may provide more insights on the relationship between the mean estimation task performance and the covariance structure of the distribution.

Following the suggestion of reviewer 5VdC, we have applied our estimator coordinate-wise to obtain a multivariate estimator and the corresponding upper bound for the multivariate case. See the discussion section for details (Proposition 6.1). The result follows from Theorem 3.5 and the union bound. Specifically the upper bound is at the order of $\sqrt{\mathrm{tr}(\Sigma)\log(nd/\delta)/n}$, which is only optimal up to logarithmic factors.

We emphasize that extending the robust estimator to the multivariate case with optimal statistical guarantees (without the logarithmic factor in the leading term) is generally a challenging problem. For example, extending the MoM estimator to the multivariate case has motivated many works; see Section 3 by Lugosi and Mendelson (2019a) for a recent review. The difficulty is due to the fact that there is no standard notion of a median for multivariate data (Small, 1990). Most of them are computationally expensive or even NP-hard. It is challenging to develop their asymptotic properties. For example, due to combinatoric nature of the median-of-means tournament (Lugosi and Mendelson, 2019b), it is very hard to write it in an M estimation framework. We will leave this to future work.

Minor comments:

> 1&2: I would suggest mentioning "heavy-tailed distribution" in either the abstract or the title. In the first sentence of the abstract, "...self-tuned robust estimators for estimating the mean of distributions with only finite variances", does "with only finite variances" refer to the robust estimator, or refer to the distribution?

Thanks for this comment. We have rewritten the first sentence of the abstract to ``In this paper, we propose self-tuned robust estimators for estimating the mean of heavy-tailed distributions, where heavy-tailed distributions refer to distributions with only finite variances."

> 3. In "the resulting estimator for the robustification parameter can adapt to the unknown variance automatically", what does the "variance" refer to?

The variance referes to the unknown data variance. We have rewritten this sentence to “......can automatically adapt to the unknown data variance and can achieve near-optimal finite-sample performance.”

Reference:

Lugosi and Mendelson (2019a), Mean estimation and regression under heavy-tailed distributions — A survey.

Lugosi and Mendelson (2019b), Sub-Gaussian estimators of the mean of a random vector.

Small (1990), A Survey of Multidimensional Medians.

Add: **Official Comment**

Replying to Response to reviewer xmsE

Official Comment by Reviewer xmsE

Official Comment Reviewer xmsE 04 Oct 2023, 16:19 Everyone

Comment:

Thanks for the authors replying to my comments.

For the new Proposition 6.1, I understand that it is an in-probability bound so the log term is necessary, but I'm wondering whether it is possible to derive the asymptotic distribution in a format similar to Theorem 4.1 and 4.2? If so, what is the asymptotic variance?

For my major concern 1 about whether this paper is suitable for ML journal or not, I would leave this up to AC for the decision.

The overall quality of this paper looks good to me. Since (1) existing minimax lower bound analysis mainly focuses on the convergence rate rather than a exact multiplicative constant, and (2) sample mean usually achieves minimax optimal, I think showing the asymptotic normality with a clear smaller variance is already sufficient.

Add: **Official Comment**

Response to reviewer xmsE

Edit

Official Comment Authors (Qiang Sun (/profile?id=~Qiang_Sun2))

05 Oct 2023, 02:23 (modified: 08 Oct 2023, 16:33) Everyone Revisions (/revisions?id=9sAixzaA6E)

Comment:

We have further proved the asymptotic result, Proposition 6.2, for the proposed multivariate estimator; see our revised manuscript. To our surprise (at least at first glance), it also achieves full asymptotic efficiency. Specifically, under certain conditions, we have:

$$\hat{\mu}(\hat{v}) - \mu^* \rightsquigarrow \mathcal{N}(0, \Sigma).$$

In hindsight, the above result is natural, as the simple coordinate-wise multivariate mean estimator also approaches to the sample mean estimator in the asymptotic limit, and thus is asymptotically efficient. This highlights another advantage of our procedure: It can easily be extended and yet still achieves full asymptotic efficiency.

Add: **Official Comment**

Replying to Response to reviewer xmsE

Official Comment by Reviewer xmsE

Official Comment   Reviewer xmsE   05 Oct 2023, 10:52   Everyone

Comment:

Thanks for your response and the additional result.

Add: **Official Comment**

Replying to Official Comment by Reviewer xmsE

Thanks

Edit

Official Comment   Authors (Qiang Sun (/profile?id=~Qiang_Sun2))   08 Oct 2023, 16:34   Everyone

Comment:

Thanks very much for your review which has helped to improve the paper significantly! If you have any other questions/confusion, please let us know.

Add: **Official Comment**

Marking this work as a long submission

Official Comment

Editors In Chief (Raia Hadsell (/profile?id=~Raia_Hadsell2), Kyunghyun Cho (/profile?id=~Kyunghyun_Cho1), Hugo Larochelle (/profile?id=~Hugo_Larochelle1), tmlr-editors@jmlr.org (/profile?id=tmlr-editors@jmlr.org), +1 more (/group/info?id=TMLR/Editors_In_Chief))

30 Aug 2023, 08:58   Authors, Editors In Chief

Comment:

Hi Qiang,

Seeing that this work has long proofs in the appendix that arguably should be read by the reviewers, we've decided to mark this submission as a long (as opposed to regular) submission. The implication is that you should expect a longer review process than for regular submissions.

Hugo

Add: **Official Comment**

No problem

Edit 🗑

Official Comment ✎ Authors (👁 Qiang Sun (/profile?id=~Qiang_Sun2)) 📅 30 Aug 2023, 13:32

👁 Authors, Editors In Chief

Comment:

Dear Hugo, Thanks! This is better as it will provide the reviewers a longer time to read it more carefully. Qiang

Add: **Official Comment**

Review Approval of Paper1503 by Action Editors

Review Approval ✎ Action Editors 📅 28 Aug 2023, 23:13 (modified: 28 Aug 2023, 23:13)

👁 Editors In Chief, Action Editors, Authors Revisions (/revisions?id=3zh7XgHxiA)

Under Review: Appropriate for Review

Comment:

Because of the technical nature of the results and the length of the proofs in the appendix, I strongly recommend to consider this paper as a "long paper" to provide enough time to the reviewers.

Add: **Official Comment**

About OpenReview (/about)
Hosting a Venue (/group?id=OpenReview.net/Support)
All Venues (/venues)

Contact (/contact)
Feedback
Sponsors (/sponsors)

Frequently Asked Questions (https://docs.openreview.net/getting-started/frequently-asked-questions)
Terms of Use (/legal/terms)
Privacy Policy (/legal/privacy)

OpenReview (/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through Code for Science & Society (https://codeforscience.org/). We gratefully acknowledge the support of the OpenReview Sponsors (/sponsors).