mathematics

MDPI

Article

Sharper Concentration Inequalities for Median-of-Mean Processes

Guangqiang Teng [1,†], Yanpeng Li [2,†], Boping Tian [1,*] and Jie Li [3,*]

[1] School of Mathematics, Harbin Institute of Technology, Harbin 150001, China; gqtenghit@163.com
[2] Department of Statistics and Data Science, National University of Singapore, 21 Lowr Kent Ridge Road, Singapore 119077, Singapore; e0914291@u.nus.edu
[3] School of Statistics, Renmin University of China, Beijing 100872, China
* Correspondence: bopingt361147@hit.edu.cn (B.T.); lijie_stat@ruc.edu.cn (J.L.)
† These authors contributed equally to this work.

Abstract: The Median-of-Mean (MoM) estimation is an efficient statistical method for handling data with contamination. In this paper, we propose a variance-dependent MoM estimation method using the tail probability of a binomial distribution. The bound of this method is better than the classical Hoeffding method under mild conditions. This method is then used to study the concentration of variance-dependent MoM empirical processes and sub-Gaussian intrinsic moment norm. Finally, we give the bound of the variance-dependent MoM estimator with distribution-free contaminated data.

Keywords: concentration inequality; Median-of-Mean; robust machine learning; contaminated data

MSC: 62B10

1. Introduction

Nowadays, there is a huge amount of data in information processing, and the data are varied. With the rapid expansion of data volume, traditional centralized data processing has gradually become unable to adapt to the current needs, which makes it possible to distribute processing power to all computers on the network.

When dealing with large amounts of data, it is inevitable to produce contaminated data which we generally call outliers. The outliers will result in low accuracy or high sensitivity of data processing tasks. Naturally, inferring probability density functions from contaminated samples is an important problem. Correspondingly, when there are no outliers in a dataset, we call such a dataset sane.

The Median-of-Mean (MoM) method is an effective way to deal with contaminated data, which divides the original data into several blocks, calculates the mean for each block, and then takes the median of these means. The literature on MoM methods can be traced back to Ref. [1]. In recent years, MoM methods have been widely used in the field of machine learning. For example, Ref. [2] used the MoM method to design estimators for kernel mean embedding and maximum mean discrepancy with excessive resistance properties to outliers; Ref. [3] applied the MoM method to achieve the optimal trade-off between accuracy and confidence under minimal assumptions in the classical statistical learning/regression problem; Ref. [4] introduced an MoM method for robust machine learning without deteriorating the estimation properties of a given estimator which is also easily computable in practice; Ref. [5] introduced a robust nonparametric density estimator combining the popular Kernel Density Estimation method and the Median-of-Means principle.

When using MoM methods to deal with contaminated data, these data often do not have obvious normal distribution characteristics but have more extensive sub-Gaussian properties; thus, non-asymptotic techniques are needed. Non-asymptotic inference can

check for updates

Citation: Teng, G.; Li, Y.; Tian, B.; Li, J. Sharper Concentration Inequalities for Median-of-Mean Processes. *Mathematics* **2023**, *11*, 3730. https://doi.org/10.3390/math11173730

Academic Editors: Huiming Zhang and Ting Yan

Received: 30 July 2023
Revised: 29 August 2023
Accepted: 29 August 2023
Published: 30 August 2023

give full play to its advantages in the case of finite samples. Especially in the field of machine learning, non-asymptotic inference can establish strict error boundaries for the desired learning program (see Ref. [6–8]). Sometimes when working with data, it is difficult to know the exact distribution; this calls for a more general study such as sub-Gaussian, sub-exponential, heavy-tailed, and bounded distributions. For example, Ref. [9] studied the non-asymptotic concentration of the heteroskedastic Wishart-type matrices; Ref. [10] constructed sub-Gaussian estimators of a mean vector under adversarial contamination and heavy-tailed data by Median-of-Mean versions of the Stahel–Donoho outlyingness and of Median Absolute Deviation functions; Ref. [11] obtained the deconvolution for some singular density errors via a combinatorial Median-of-Mean approach and assessed the estimator quality by establishing non-asymptotic risk bounds.

To obtain a clear picture of robust estimation from a non-asymptotic viewpoint, variance-dependent MoM methods based on binomial tail probability are mainly studied, including uncontaminated and contaminated cases. The paper proceeds as follows. We first provide a variance-dependent MoM-estimator bias inequality by using bounds on binomial tails with unbounded samples, whose bias bound is tighter than the classical Hoeffding's bound (see Section 2). Then, by the variance-dependent MoM inequality, we obtain the generalization bound via entropic complexity (see Section 3.1) and the non-asymptotic property via Sub-Gaussian intrinsic moment norm (see Section 3.2). Finally, the variance-dependent MoM inequality with contamination data is illustrated in Section 4.

2. Variance-Dependent Median-of-Mean Estimator without Outliers

The MoM method was originally introduced on page 242 of Ref. [1]; it reinforces the effect of the empirical mean on the heavy-tail distribution while inheriting its efficiency on the light-tail distribution. The MoM estimator is derived as follows.

Without loss of generality, suppose that the sample data $X_1, X_2, \ldots, X_n$ are decomposed into K blocks, with each block including B observations, that is to say, $n = KB$. We first compute the mean of each block, which leads to estimators $\hat{\mu}_1, \cdots, \hat{\mu}_K$ and each estimator is based on B observations. Then, the MoM estimator is given by the median of all these estimators, i.e.,

$$\mathrm{MoM}_K[\mu] = \mathrm{median}(\hat{\mu}_1, \cdots, \hat{\mu}_K).$$

It turns out that, even with a very mild condition $\mathrm{Var}(X) = \sigma^2 < \infty$, the MoM estimator has a nice concentration inequality under finite sample case.

Given the i.i.d. sample $X_1, X_2, \ldots, X_n$ with mean μ_0 and finite variance σ^2, using Hoeffding's inequality, Proposition 1 in Ref. [12] produces the following concentration inequality:

$$\mathbb{P}(|\mathrm{MoM}_K[\mu] - \mu_0| > t) \leq \exp\left(-\frac{nt^2}{27\sigma^2}\right)$$

where $t = \sigma\sqrt{(2+\delta)/B} \geq \sigma\sqrt{\frac{2\sqrt{\pi}-\sqrt{2}}{B}}$—see detailed description in Remark 1.

When additional conditions are applied to the distribution under consideration, stricter boundaries can be obtained, such as our results on binomial tails (Theorem 1), which can be better.

In fact, sometimes we need to block the data, but the minimum number of samples per block is often a concern, because it involves efficiency and robustness issues, and, from a statistical point of view, the effect of variance is taken into account. The following theorem takes into account the partitioning of variance effects and yields the variance-dependent MoM inequality.

Theorem 1. *Given the i.i.d. samples $X_1, X_2, \ldots, X_n$ with mean μ_0 and finite variance σ^2, for $\forall \delta \geq \frac{\sqrt{2}}{\sqrt{\pi}-\sqrt{2}}$, there exists $B \in \mathbb{N}$ and $\varepsilon > 0$, such that $B\varepsilon^2 \geq (2+\delta)\sigma^2$. Then, the MoM estimator has the following concentration inequality:*

$$\mathbb{P}(|\mathrm{MoM}_K[\mu] - \mu_0| > t) \leq \exp\left(-\frac{0.0976nt^2}{\sigma^2}\right) \tag{1}$$

where $t = \sigma\sqrt{(2+\delta)/B}$.

A powerful feature of Theorem 1 is that X_is can be unbounded in this case. In addition, finite sample exponential concentration is not easy to obtain if only variance exists (see Ref. [13]). And Theorem 1 provides the basis for further obtaining the inequality with outliers. In the process of proving the theorem, we used the following lemma.

Lemma 1 (Theorem 1 of [14]). *Suppose* $S_n \sim \mathrm{Bin}(n, p), a > p \in (0,1)$, *and* $1 \leqslant an \leqslant n-1$. *If* $an \in \mathbb{N}$, *then*

$$\mathbb{P}(S_n \geqslant an) \leqslant \frac{1}{1-r}\frac{1}{\sqrt{2\pi a(1-a)n}}e^{-nD(a\|p)}.$$

where $r = r(a,p) := \frac{p(1-a)}{a(1-p)}$, *and* $D(a\|p) := a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}$ *is the KL divergence between Bernoulli distributions with parameters a and p. If* $an \notin \mathbb{N}$, *the bound still holds, but it can be tightened by replacing a with* $a^* := \lceil an \rceil / n$.

Now, we give a detailed proof of Theorem 1.

Proof of Theorem 1. First, observe that the event

$$\{|\mathrm{MoM}_K[\mu] - \mu_0| > \epsilon\} \text{ for } \forall \epsilon \geq 0$$

implies that at least $K/2$ of $\hat{\mu}_\ell (\ell = 1, \ldots, K)$ has to be outside ϵ distance to μ_0 for $\forall \epsilon \geq 0$. Namely,

$$\{|\mathrm{MoM}_K[\mu] - \mu_0| > \epsilon\} \subset \left\{\sum_{\ell=1}^{K} \mathbf{1}(|\hat{\mu}_\ell - \mu_0| > \epsilon) \geq \frac{K}{2}\right\} \text{ for } \forall \epsilon \geq 0.$$

Here, it is assumed that K is an even number. When K is an odd number, take at least $\lceil K/2 \rceil$, and the same can be said. For the convenience of writing, the following process of proof only writes the case of at least $K/2$, while proving the case of $\lceil K/2 \rceil$ is no difference.

Define $Z_\ell = \mathbf{1}(|\hat{\mu}_\ell - \mu_0| > \epsilon)$ and let $\bar{p} := \tilde{p}_{\epsilon,B} = \mathbb{E}(Z_\ell) = \mathbb{P}(|\hat{\mu}_\ell - \mu_0| > \epsilon)$. Note the theorem condition and the Chebyshev's inequality (see p. 239 in Ref. [15]), which imply that there exits $B \in \mathbb{N}$ and $\varepsilon > 0$ such that

$$\tilde{p} := \tilde{p}_{\varepsilon,B} = \mathbb{P}(|\hat{\mu}_\ell - \mu_0| > \varepsilon) \leq \frac{\sigma^2}{B\varepsilon^2} < \frac{1}{2}. \tag{2}$$

In fact, the detailed derivation process is as follows:

$$\begin{aligned}
\mathbb{P}(|\hat{\mu}_\ell - \mu_0| > \varepsilon) &\leq \frac{\mathrm{Var}(\hat{\mu}_\ell)}{\varepsilon^2} \\
&= \frac{\mathrm{Var}\left(\frac{X_{l1}+\cdots+X_{lB}}{B}\right)}{\varepsilon^2} \\
&= \frac{\frac{1}{B^2}\mathrm{Var}\left(\sum_{i=1}^{B} X_{li}\right)}{\varepsilon^2} \\
&= \frac{\frac{1}{B^2}\sum_{i=1}^{B}\mathrm{Var}(X_{li})}{\varepsilon^2} \\
&= \frac{\frac{1}{B^2}B\sigma^2}{\varepsilon^2} \\
&= \frac{\sigma^2}{B\varepsilon^2}.
\end{aligned}$$

The random variables $Z_\ell \sim \text{Bernoulli}(\tilde{p})$ are i.i.d. because of the i.i.d. samples $X_1, X_2, \cdots, X_n$. Applying Lemma 1 (with $a = 1/2$, $n = K$, and $p = \tilde{p}$ in Lemma 1) to the summations gives

$$\mathbb{P}(|\text{MoM}_K[\mu] - \mu_0| > \varepsilon) \leq \mathbb{P}\left(\sum_{\ell=1}^{K} Z_\ell \geq \frac{K}{2}\right) \leq \frac{1-\tilde{p}}{1-2\tilde{p}}\sqrt{\frac{2}{\pi K}}e^{-KD\left(\frac{1}{2}||\tilde{p}\right)},$$

where $D\left(\frac{1}{2}||\tilde{p}\right) = \frac{1}{2}\log\left(\frac{1}{4\tilde{p}(1-\tilde{p})}\right)$.

Setting $B \geq (2+\delta)\sigma^2/\varepsilon^2 > 2\sigma^2/\varepsilon^2$ for $\forall \delta > 0$ satisfies Equation (2); then,

$$\begin{aligned}\mathbb{P}\left(|\text{MoM}_K[\mu] - \mu_0| > \sigma\sqrt{\frac{(2+\delta)K}{n}}\right) &\leq \frac{1-\tilde{p}}{1-2\tilde{p}}\sqrt{\frac{2}{\pi K}}e^{-KD\left(\frac{1}{2}||\tilde{p}\right)} \\ &= \left(1+\frac{\tilde{p}}{1-2\tilde{p}}\right)\sqrt{\frac{2}{\pi K}}e^{-KD\left(\frac{1}{2}||\tilde{p}\right)} \\ &\leq \frac{\delta+1}{\delta}\sqrt{\frac{2}{\pi K}}\left(1+\frac{\delta^2}{4+4\delta}\right)^{-\frac{K}{2}}\end{aligned}$$

When $K = 1$, we set $\delta \geq \frac{\sqrt{2}}{\sqrt{\pi}-\sqrt{2}} \approx 3.95$ so that $\frac{\delta+1}{\delta}\sqrt{\frac{2}{\pi K}} \leq \frac{\delta+1}{\delta}\sqrt{\frac{2}{\pi}} \leq 1 (K = 1, \cdots, n)$. Then, it follows that

$$\mathbb{P}\left(|\text{MoM}_K[\mu] - \mu_0| > \sigma\sqrt{\frac{(2+\delta)K}{n}}\right) \leq \left(1+\frac{\delta^2}{4+4\delta}\right)^{-\frac{K}{2}}$$

for $1 \leq K \leq n$ and $\delta \geq \sqrt{2}/(\sqrt{\pi}-\sqrt{2})$.

Now, taking $t := \sigma\sqrt{(2+\delta)K/n}$ gives

$$\mathbb{P}(|\text{MoM}_K[\mu] - \mu_0| > t) \leq \exp\left(-\frac{nt^2}{2(2+\delta)\sigma^2}\ln\left(1+\frac{\delta^2}{4+4\delta}\right)\right)$$

The function $g(\delta) = -\frac{1}{2+\delta}\ln\left(1+\frac{\delta^2}{4+4\delta}\right)(\delta \geq \frac{\sqrt{2}}{\sqrt{\pi}-\sqrt{2}})$ is a monotonically decreasing function, so its maximum is $g\left(\sqrt{2}/(\sqrt{\pi}-\sqrt{2})\right) \approx -0.0976$.

This then leads to the final result:

$$\mathbb{P}(|\text{MoM}_K[\mu] - \mu_0| > t) \leq \exp\left(-\frac{0.0976nt^2}{\sigma^2}\right).$$

□

Remark 1. *The classical result by Hoeffding inequality shows that (see Proposition 1 in Ref.* [12]*)*

$$\mathbb{P}\left(|\text{MoM}_K[\mu] - \mu_0| > \sigma\sqrt{(2+\delta)K/n}\right) \leq e^{-K\frac{\delta^2}{2(2+\delta)^2}}$$

Similarily, to obtain a sharp constant, one can consider $t := \sigma\sqrt{(2+\delta)K/n}$*; then,*

$$\mathbb{P}(|\text{MoM}_K[\mu] - \mu_0| > t) \leq \exp\left(-\frac{nt^2\delta^2}{2\sigma^2(2+\delta)^3}\right)$$

and the function

$$g(\delta) = \frac{\delta^2}{(2+\delta)^3}$$

achieve the unique maximum point at $\delta = 4$ *with* $g(4) = 2/27$. *It follows that*

$$\mathbb{P}(|\mathrm{MoM}_K[\mu] - \mu_0| > t) \leq \exp\left(-\frac{nt^2}{27\sigma^2}\right).$$

Remark 2. *The efficient interval of* t *is an interesting issue. By the construction of* $t = \sigma\sqrt{(2+\delta)K/n}$, *it follows that* $\sqrt{(2+\delta)/n} \leq t/\sigma \leq \sqrt{2+\delta}$ *since* $1 \leq K \leq n$.

Remark 3. *In Theorem 1, we substitute* $t = \sigma\sqrt{(2+\delta)/B}$ *into inequality (1) to produce*

$$\mathbb{P}\left(|\mathrm{MoM}_K[\mu] - \mu_0| > \sigma\sqrt{\frac{2+\delta}{B}}\right) \leq \exp\left(-\frac{0.0976n\sigma^2(2+\delta)}{B\sigma^2}\right)$$

Since $\delta \geq \frac{\sqrt{2}}{\sqrt{\pi}-\sqrt{2}} \approx 3.95 > 2$, *we have*

$$\mathbb{P}\left(|\mathrm{MoM}_K[\mu] - \mu_0| > 2\sigma\sqrt{\frac{K}{n}}\right) \leq e^{-0.5807K}.$$

This result is better than the bound $e^{-K/8}$ *of level-dependent sub-Gaussian estimators. Of course, our conditions are more stringent (see Proposition 12 in Ref.* [16]*).*

3. Applications

In this section, we use the proposed sharper concentration inequalities for MoM estimators to perform two applications in statistical machine learning.

3.1. Concentration for Supremum of Variance-Dependent MoM Empirical Processes

Let $\psi(x) \in \mathcal{B}_L$ and $|\psi(x)| \leq M_0 < \infty$, where $\mathcal{B}_L$ is a ball of the Lipschitz functions space and M_0 is a constant. Let $\mathrm{P}\psi = \mathbb{E}\psi = \int \psi dP$.

To derive the concentration inequality for the supremum of variance-dependent MoM empirical processes, the following auxiliary Lemma 2 is necessary, whose proof is trivial and thus omitted.

Lemma 2. $|\operatorname{med}(a) - \operatorname{med}(b)| \leq \|a - b\|_\infty$ *for* $a, b \in B_{\mathcal{L}}$ *where* $\operatorname{med}(a)$ *means the value of the function* $a(x)$ *at the midpoint of the domain, and the same is true for* $\operatorname{med}(b)$.

By Lemma 2, for $\forall \phi \in \mathcal{B}_L$, we have

$$\begin{aligned} |\mathrm{MoM}_K[\phi] - \mathrm{P}\phi| &\leq |\mathrm{MoM}_K[\phi] - \mathrm{MoM}_K[\psi]| + |\mathrm{P}(\phi - \psi)| + |\mathrm{MoM}_K[\psi] - \mathrm{P}\psi| \\ &\leq \|\phi - \psi\|_\infty + \|\phi - \psi\|_\infty + |\mathrm{MoM}_K[\psi] - \mathrm{P}\psi| \\ &= 2\|\phi - \psi\|_\infty + |\mathrm{MoM}_K[\psi] - \mathrm{P}\psi| \end{aligned} \tag{3}$$

Let $\psi_1, \cdots, \psi_{\mathcal{N}(\xi, \mathcal{B}_L, \|\cdot\|_\infty)}$ be a ξ-covering of $\mathcal{B}_L$ w.r.t. $\|\cdot\|_\infty$. It is well-known that there exist constants $C_L > 0$ and $r \geq 1$, such that

$$\log(\mathcal{N}(\xi, \mathcal{B}_L, \|\cdot\|_\infty)) \leq C_L\left(\frac{1}{\xi}\right)^r, \forall \xi > 0 \tag{4}$$

where $\mathcal{N}(\xi, \mathcal{B}_L, \|\cdot\|_\infty)$ denotes the number of $\|\cdot\|_\infty$-balls of radius $\xi > 0$ needed to cover class $\mathcal{B}_L$, and C_L is a universal constant depending only on $\mathcal{B}_L$.

Put $\mathcal{N} = \mathcal{N}(\xi, \mathcal{B}_L, \|\cdot\|_\infty)$ for simplicity. By definition of $\mathcal{N}$, for $\forall i \in \{1, \cdots, \mathcal{N}\}$, s.t.

$$\|\phi - \psi_i\|_\infty \leq \xi$$

Then, (3) becomes

$$|\mathrm{MoM}_K[\phi] - \mathrm{P}\phi| \le 2\tilde{\xi} + |\mathrm{MoM}_K[\psi_i] - \mathrm{P}\psi_i| \tag{5}$$

Then, by Theorem 1, the union bound for $\{\psi_i\}_{i=1}^{\mathcal{N}}$ gives that

$$\mathbb{P}\left(\max_{1\le i\le \mathcal{N}} |\mathrm{MoM}_K[\psi_i] - \mathrm{P}\psi_i| \le \sigma\sqrt{\frac{-\ln\delta}{0.0976\mathcal{N}}}\right) \ge 1-\delta. \tag{6}$$

Together, (4)–(6) give

$$\mathbb{P}\left(\sup_{\phi\in\mathcal{B}_L} |\mathrm{MoM}_K[\phi] - \mathrm{P}\phi| \le 2\xi + \sigma\sqrt{\frac{-\ln\delta}{0.0976\mathcal{N}}}\right) \ge 1-\delta.$$

Put $\xi = \sqrt{\frac{C_L}{\mathcal{N}\xi}}$, i.e., $\xi = \left(\frac{C_L}{\mathcal{N}}\right)^{\frac{1}{r+2}}$; then, for $\forall\phi \in \mathcal{B}_L$ and $\delta \in (0,1)$, we have

$$\mathbb{P}\left(\sup_{\phi\in\mathcal{B}_L} |\mathrm{MoM}_K[\phi] - \mathrm{P}\phi| \le 2\left(\frac{C_L}{\mathcal{N}}\right)^{\frac{1}{r+2}} + \sigma\sqrt{\frac{-\ln\delta}{0.0976\mathcal{N}}}\right) \ge 1-\delta.$$

3.2. Concentration for Variance-Dependent MoM Intrinsic Moment Norm

A centered random variable X is called sub-Gaussian if

$$\mathrm{E}e^{sX} \le e^{s^2\sigma_G^2/2} \text{ for } \forall\, s \in \mathbb{R},$$

where the quantity $\sigma_G > 0$ is named as the sub-Gaussian parameter. In non-asymptotic statistics, because the collected sub-Gaussian data is often unstable, sometimes it is not possible to directly use the empirical moment-generating function to estimate the sub-Gaussian parameter such as variance-type parameters of sub-Gaussian distributions (see Ref. [17]). This requires us to use the sub-Gaussian intrinsic moment norm for estimation. The definition of intrinsic moment norm is as follows.

Definition 1 (Intrinsic moment norm, see Definition 2 in Ref. [17])**.** *The sub-Gaussian intrinsic moment norm is defined as*

$$\|X\|_G := \max_{k\ge 1}\left[\frac{2^k k!}{(2k)!}\mathrm{E}X^{2k}\right]^{1/(2k)} = \max_{k\ge 1}\left[\frac{1}{(2k-1)!!}\mathrm{E}X^{2k}\right]^{1/(2k)},$$

where $n!! = \prod_{j=0}^{\left[\frac{n}{2}\right]-1}(n-2j) = n(n-2)(n-4)\cdots$ *for* $n \in \mathbb{N}$.

As the amount of computation increases, so does the importance of the distributed MoM approach, with the corresponding intrinsic moment norm estimator defined below.

Definition 2 (see Equation (7) in Ref. [17])**.** *Let* $[K] = \{1,\cdots,K\}$ *and* B_s *be the number of samples in the s-th block. The MOM estimator for sub-Gaussian intrinsic moment norm is given by*

$$\widehat{\|X\|}_{b,G} := \max_{1\le k\le\kappa_n} \underset{s\in[K]}{\mathrm{median}}\left\{\left[[(2k-1)!!]^{-1}\mathrm{P}_B^{B_s}X^{2k}\right]^{1/(2k)}\right\}$$

where $\mathrm{P}_B^{B_s}X = B^{-1}\sum_{i\in B_s}X_i (s = 1,\cdots,K)$.

Definition 3. *For any* $B \in \mathbb{N}$ *and* $1 \le k \le \kappa_n$,

$$\bar{g}_{k,B}(\sigma_k) := 1 - \left[\mathrm{E}\,X^{2k}/(2k-1)!!\right]^{-\frac{1}{2k}} \max_{1\le j\le\kappa_n}\left[-2B^{-\frac{1}{2}}\sigma_j^j/(\mathrm{E}\,X^{2j}) + \mathrm{E}\,X^{2j}/(2j-1)!!\right]^{\frac{1}{2j}}$$

and $\underline{g}_{k,B}(\sigma_k) := \left[2B^{-1/2}\sigma_k^k/(\mathrm{E}\,X^{2k}) + 1\right]^{1/(2k)} - 1.$

Theorem 2. *Suppose, for* $\forall \varepsilon > 0$ *and* $\forall n \in \mathbb{N}$, *there exits* $B \in \mathbb{N}$, *such that* $\sqrt{\mathrm{Var}X^{2k}} < \varepsilon\sqrt{B/2} \le \sigma_k^k$ *where* $\{\sigma_k\}_{k=1}^{\kappa_n}$ *is a finite constant sequence. Then, we have*

$$\mathbb{P}\left\{\|X\|_G \le \left[1 - \max_{1\le k\le\kappa_n} \bar{g}_{k,B}(\sigma_k)\right]^{-1} \widehat{\|X\|}_{b,G}\right\} > 1 - \kappa_n e^{-0.3904K}$$

and

$$\mathbb{P}\left\{\|X\|_G > \left[1 + \max_{1\le k\le\kappa_n} \underline{g}_{k,B}(\sigma_k)\right]^{-1} \widehat{\|X\|}_{b,G}\right\} > 1 - \kappa_n e^{-0.3904K}.$$

Remark 4. *Let* $K = n/B$; *we then obtain distributed samples that satisfy Theorem 2.*

Remark 5. *The key coefficient* $-0.3904 < -0.125$. *In fact, the key coefficient of Theorem 3 in Ref.* [17] *without outliers is* -0.125, *as long as* $\eta(\varepsilon) = 1$ *is taken. This means that our boundary is better than the boundary in Ref.* [17].

Proof of Theorem 2. From Definitions 1 and 2, we have

$$\|X\|_G = \max_{1\le k\le\kappa_n}\left[\frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)} \tag{7}$$

and

$$\widehat{\|X\|}_{b,G} = \max_{1\le k\le\kappa_n} \mathrm{median}_{s\in[K]}\left\{\left[\frac{1}{(2k-1)!!}\cdot \mathrm{P}_B^{B_s} X^{2k}\right]^{1/(2k)}\right\}. \tag{8}$$

Recall that $\underline{g}_{k,B}(\sigma_k)$ and $\bar{g}_{k,B}(\sigma_k)$ are the sequences s.t.

$$\begin{aligned}&\left[\mathrm{E}X^{2k}/(2k-1)!!\right]^{1/(2k)}(1-\bar{g}_{k,B}(\sigma_k))\\ &= \max_{1\le k\le\kappa_n}\left[-2B^{-1/2}\sigma_k^k/\left(\mathrm{E}X^{2k}\right) + \mathrm{E}X^{2k}/(2k-1)!!\right]^{1/(2k)}\end{aligned} \tag{9}$$

and

$$\left[2B^{-1/2}\sigma_k^k/\left(\mathrm{E}X^{2k}\right) + 1\right]^{1/(2k)} = 1 + \underline{g}_{k,B}(\sigma_k) \tag{10}$$

for any $B \in \mathbb{N}$ and $1 \le k \le \kappa_n$.

For the first inequality of Theorem 2, we have, by (7),

$$\begin{aligned}&\mathbb{P}\left\{\widehat{\|X\|}_{b,G} \le \left[1 - \max_{1\le k\le\kappa_n}\bar{g}_{k,B}(\sigma_k)\right]\|X\|_G\right\}\\ &= \mathbb{P}\left\{\widehat{\|X\|}_{b,G} \le \max_{1\le k\le\kappa_n}\left[\frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)}\left(1 - \max_{1\le k\le\kappa_n}\bar{g}_{k,B}(\sigma_k)\right)\right\}\\ &\le \mathbb{P}\left\{\widehat{\|X\|}_{b,G} \le \max_{1\le k\le\kappa_n}\left[\frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)}(1-\bar{g}_{k,B}(\sigma_k))\right\}\\ [\text{By (9)}]\ &= \mathbb{P}\left\{\widehat{\|X\|}_{b,G} \le \left[-\frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} + \frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)}\right\}\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \operatorname{median}_{s\in[K]}\left\{ \left[\frac{1}{(2k-1)!!}\cdot \mathrm{P}_B^{B_s} X^{2k}\right]^{1/(2k)} \right\} \leq \right. \\
&\qquad \left. \left[-\frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} + \frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)} \right\} \\
&= \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \operatorname{median}_{s\in[K]}\left\{ \frac{1}{(2k-1)!!}\cdot \mathrm{P}_B^{B_s} X^{2k}\right\} \leq \right. \\
&\qquad \left. \frac{\mathrm{E}X^{2k}}{(2k-1)!!} - \frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} \right\} \\
&= \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \operatorname{median}_{s\in[K]}\left\{ \frac{1}{(2k-1)!!}\cdot \left[\mathrm{P}_B^{B_s} X^{2k} - \mathrm{E}X^{2k}\right]\right\} \leq \right. \\
&\qquad \left. -\frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} \right\} \\
&< \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \left|\operatorname{median}_{s\in[K]}\left\{ \mathrm{P}_B^{B_s}\left[X^{2k} - \mathrm{E}^{2k}\right]\right\}\right| \geq \sigma_k^k\cdot\frac{2}{B^{1/2}} \right\} \\
&\leq \kappa_n e^{-0.3904K},
\end{aligned}
$$

where the last inequality is by Theorem 1 and the assumption in Theorem 2.

Let $\underline{g}_B(\sigma) := \max_{1\leq k\leq\kappa_n} \underline{g}_{k,B}(\sigma_k)$. For the second inequality of Theorem 2, the definition of $\underline{g}_{k,B}(\sigma_k)$ implies

$$
\begin{aligned}
&\mathbb{P}\left\{ \|X\|_G \leq \frac{\widehat{\|X\|}_{b,G}}{1+\underline{g}_B(\sigma)} \right\} \\
&= \mathbb{P}\left\{ \widehat{\|X\|}_{b,G} \geq \max_{1\leq k\leq\kappa_n}\left[\frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} + \frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)} \right\} \\
&\leq \mathbb{P}\left\{ \max_{1\leq k\leq\kappa_n} \operatorname{median}_{s\in[K]}\left\{ \left[\frac{1}{(2k-1)!!}\cdot \mathrm{P}_B^{B_s} X^{2k}\right]^{1/(2k)} \right\} \geq \right. \\
&\qquad \left. \left[\frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} + \frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)} \right\} \\
&\leq \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \operatorname{median}_{s\in[K]}\left\{ \left[\frac{1}{(2k-1)!!}\cdot \mathrm{P}_B^{B_s} X^{2k}\right]^{1/(2k)} \right\} \geq \right. \\
&\qquad \left. \left[\frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} + \frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)} \right\} \\
&= \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \operatorname{median}_{s\in[K]}\left\{ \frac{1}{(2k-1)!!}\cdot \mathrm{P}_B^{B_s} X^{2k}\right\} \geq \left[\frac{\sigma_k^k}{(2k-1)!!}\cdot\frac{2}{B^{1/2}} + \frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right] \right\} \\
&= \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \operatorname{median}_{s\in[K]}\left\{ \mathrm{P}_B^{B_s} X^{2k}\right\} \geq \left[\frac{2\sigma_k^k}{B^{1/2}} + \mathrm{E}X^{2k}\right] \right\} \\
&= \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \operatorname{median}_{s\in[K]}\left\{ \mathrm{P}_B^{B_s}\left[X^{2k} - \mathrm{E}X^{2k}\right]\right\} \geq \frac{2\sigma_k^k}{B^{1/2}} \right\} \\
&< \sum_{k=1}^{\kappa_n} \mathbb{P}\left\{ \left|\operatorname{median}_{s\in[K]}\left\{ \mathrm{P}_B^{B_s}\left[X^{2k} - \mathrm{E}X^{2k}\right]\right\}\right| \geq \frac{2\sigma_k^k}{B^{1/2}} \right\} \\
&\leq \kappa_n e^{-0.3904K},
\end{aligned}
$$

where the last inequality is by Theorem 1 and the assumption in Theorem 2. □

4. Concentration for Variance-Dependent MoM with Distribution-Free Outliers

In the field of big data and artificial intelligence, most work involves dealing with abnormal data. Sometimes we cannot find each outlier directly, but we can obtain a rough idea of the total number of outliers. For example, sometimes there may be abnormal economic activities in a certain region, but the specific company or person who is abnormal may not be known for the time being; however, the total number of companies and the total population in the region are still known.

Based on such information, how to accurately estimate the characteristics of all samples containing outliers is an important problem. In this section, we introduce the concept of variance-dependent MoM estimator with outliers as the following theorem.

Theorem 3. *Suppose that*
(H.1) Sample $[n] = \{X_1, X_2, \ldots, X_n\}$ *contains* $n - n_{\mathcal{O}}$ *i.i.d. inliers with finite mean* μ_0 *and finite variance* σ^2. *And* $n_{\mathcal{O}}$ *outliers, upon which no assumption is made.*
(H.2) Set $K = K_{\mathcal{O}} + K_{\mathcal{S}}$, *where* $K_{\mathcal{O}}$ *is the number of blocks containing at least one outlier and* $K_{\mathcal{S}}$ *is the number of sane blocks containing no outlier. For* $\forall t > 0$, *there exists a function* $\eta(\varepsilon_{\mathcal{O}}) \in (1/2, 1)$, *such that* $K \geq \max\left(2, \left\lceil \frac{1}{2\eta(\varepsilon_{\mathcal{O}})-1} \right\rceil, \left\lceil \frac{(2\eta(\varepsilon_{\mathcal{O}})-1)nt^2}{2\eta(\varepsilon_{\mathcal{O}})\sigma^2} \right\rceil\right)$ *and* $K_{\mathcal{S}} \geq \eta(\varepsilon_{\mathcal{O}})K$, *where* $\varepsilon_{\mathcal{O}} := n_{\mathcal{O}}/n$.

Then, for $\forall t > 0$, *we have*

$$\mathbb{P}\{|\mathrm{MoM}_K[\mu] - \mu_0| \leq t\} \geq 1 - \exp\left(-\left(\frac{(2\eta(\varepsilon_{\mathcal{O}})-1)nt^2}{2\eta(\varepsilon_{\mathcal{O}})\sigma^2} - 1\right)\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{(2\eta(\varepsilon_{\mathcal{O}})-1)}{2\eta(\varepsilon_{\mathcal{O}})}\right).$$

Remark 6. *For the number* $n_{\mathcal{O}}$ *and* $K_{\mathcal{O}}$, *when one divides n samples evenly into K blocks, an extreme case is to assume that the blocks that do not conform to one's preferences are full of outliers, such as* $K_{\mathcal{O}}$ *blocks, and the blocks that conform to one's preferences have no outliers, such as* $K_{\mathcal{S}}$ *blocks; then, one has* $\varepsilon_{\mathcal{O}} = n_{\mathcal{O}}/n = K_{\mathcal{O}}/K$.

Remark 7. *For the function* $\eta(\varepsilon_{\mathcal{O}})$, *we can write a concrete expression to show that such a function exists, for example,* $\eta(\varepsilon_{\mathcal{O}}) = (1 + 2\varepsilon_{\mathcal{O}})/2 \in (1/2, 1)$, *where* $\varepsilon_{\mathcal{O}} \in (0, 1/2)$. *But there must be more than one expression, so the non-concrete function* $\eta(\varepsilon_{\mathcal{O}})$ *is more appropriate for this theorem.*

In fact, there is an adaptive way to generate block number K, but we do not show the specific calculation here; see Ref. [18] for more detail. Now, we give a detailed proof of Theorem 3.

Proof of Theorem 3. In the sane blocks, in the number of blocks whose sample mean is no more than t from the population mean μ_0 is at least $K/2$, the distance between the population MoM and the population mean μ_0 is no more than t, which is mathematically expressed as follows: for $\forall\, t > 0$, we have

$$\begin{aligned}\{|\mathrm{MoM}_K[\mu] - \mu_0| \leq t\} &\supset \left\{\sum_{i\in[K_{\mathcal{S}}]} \mathbf{1}_{\{|\hat{\mu}_i - \mu_0| \leq t\}} \geq \frac{K}{2}\right\} \\ &\supset \left\{\sum_{i\in[K_{\mathcal{S}}]} \mathbf{1}_{\{|\hat{\mu}_i - \mu_0| \leq t\}} \geq \frac{K_{\mathcal{S}}}{2\eta(\varepsilon_{\mathcal{O}})}\right\}.\end{aligned}$$

Further, the following formula is established:

$$\mathbb{P}\{|\mathrm{MoM}_K[\mu]-\mu_0|\le t\}\ge\mathbb{P}\left\{\sum_{i\in[K_{\mathcal{S}}]}\mathbf{1}_{\{|\hat{\mu}_i-\mu_0|\le t\}}\ge\frac{K_{\mathcal{S}}}{2\eta(\varepsilon_{\mathcal{O}})}\right\}. \tag{11}$$

From the condition (H.4), we have $1\le\frac{K_{\mathcal{S}}}{2\eta(\varepsilon_{\mathcal{O}})}\le K_{\mathcal{S}}-1$ and

$$K-1\ge K_{\mathcal{S}}\ge\eta(\varepsilon_{\mathcal{O}})K\ge 2\eta(\varepsilon_{\mathcal{O}})\ge 1+\frac{1}{K_{\mathcal{S}}-1}>1\ \text{ when }\ K\ge 2. \tag{12}$$

Applying Theorem 2 in Ref. [14], we can obtain the lower bound of Formula (11), i.e.,

$$\begin{aligned}&\mathbb{P}\left\{\sum_{i\in[K_{\mathcal{S}}]}\mathbf{1}_{\{|\hat{\mu}_i-\mu_0|\le t\}}\ge\frac{K_{\mathcal{S}}}{2\eta(\varepsilon_{\mathcal{O}})}\right\}\\&\ge\frac{1-\frac{c}{K_{\mathcal{S}}}}{1-r}\eta(\varepsilon_{\mathcal{O}})\sqrt{\frac{2}{\pi K_{\mathcal{S}}(2\eta(\varepsilon_{\mathcal{O}})-1)}}e^{-K_{\mathcal{S}}D\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\middle|\middle|\tilde{p}_{\mathcal{S}}\right)}\end{aligned} \tag{13}$$

where $c=c(r)=\frac{4\eta^2(\varepsilon_{\mathcal{O}})}{2\eta(\varepsilon_{\mathcal{O}})-1}\left[1+\frac{r(1+r)}{(1-r)^2}\right]$, $r=r\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})},\tilde{p}_{\mathcal{S}}\right)=\frac{\tilde{p}_{\mathcal{S}}(2\eta(\varepsilon_{\mathcal{O}})-1)}{1-\tilde{p}_{\mathcal{S}}}$ and

$$D\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\middle|\middle|\tilde{p}_{\mathcal{S}}\right)=\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{1}{2\eta(\varepsilon_{\mathcal{O}})\tilde{p}_{\mathcal{S}}}+\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})(1-\tilde{p}_{\mathcal{S}})}.$$

On the other hand, by Chebyshev's inequality (see p. 239 in Ref. [15]), we have

$$1-\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}<1-\tilde{p}_{\mathcal{S}}=\mathbb{P}(|\hat{\mu}_i-\mu_0|>t)\le\frac{\sigma^2}{Bt^2}=\frac{K\sigma^2}{nt^2}\le 1\ \text{ for }\ \forall t>0\ \ (i=1,\cdots,K_{\mathcal{S}}). \tag{14}$$

Thus, $\tilde{p}_{\mathcal{S}}\in[1-\frac{K\sigma^2}{nt^2},\frac{1}{2\eta(\varepsilon_{\mathcal{O}})})$ and $r\in\left[\frac{(nt^2-K\sigma^2)(2\eta(\varepsilon_{\mathcal{O}})-1)}{K\sigma^2},1\right)$. Because of $\eta(\varepsilon_{\mathcal{O}})K\le K_{\mathcal{S}}\le K-1$ and $\eta(\varepsilon_{\mathcal{O}})\in(1/2,1)$, the inequality (13) can be written as

$$\mathbb{P}\left\{\sum_{i\in[K_{\mathcal{S}}]}\mathbf{1}_{\{|\hat{\mu}_i-\mu_0|\le t\}}\ge\frac{K_{\mathcal{S}}}{2\eta(\varepsilon_{\mathcal{O}})}\right\}\ge 1-e^{-K_{\mathcal{S}}D\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\middle|\middle|\tilde{p}_{\mathcal{S}}\right)}\ge 1-e^{-(K-1)D\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\middle|\middle|\tilde{p}_{\mathcal{S}}\right)} \tag{15}$$

where

$$1+\frac{1-\frac{c}{K_{\mathcal{S}}}}{1-r}\eta(\varepsilon_{\mathcal{O}})\sqrt{\frac{2}{\pi K_{\mathcal{S}}(2\eta(\varepsilon_{\mathcal{O}})-1)}}\ge e^{K_{\mathcal{S}}D\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\middle|\middle|\tilde{p}_{\mathcal{S}}\right)}. \tag{16}$$

The inequality (16) can be valid, for example, if $\eta(\varepsilon_{\mathcal{O}})$ is infinitely close to $1/2$.

From $\tilde{p}_{\mathcal{S}}\in[1-\frac{K\sigma^2}{nt^2},\frac{1}{2\eta(\varepsilon_{\mathcal{O}})})$, we have the minimum bound of $D\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\middle|\middle|\tilde{p}_{\mathcal{S}}\right)$, i.e.,

$$\begin{aligned}D\left(\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\middle|\middle|\tilde{p}_{\mathcal{S}}\right)&>\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{1}{2\eta(\varepsilon_{\mathcal{O}})\frac{1}{2\eta(\varepsilon_{\mathcal{O}})}}+\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})(1-1+\frac{K\sigma^2}{nt^2})}\\&=\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{(2\eta(\varepsilon_{\mathcal{O}})-1)nt^2}{2\eta(\varepsilon_{\mathcal{O}})K\sigma^2}.\end{aligned} \tag{17}$$

Substituting Equation (17) into Equation (15), we have

$$\begin{aligned}&\mathbb{P}\left\{\sum_{i\in[K_{\mathcal{S}}]}\mathbf{1}_{\{|\hat{\mu}_i-\mu_0|\le t\}}\ge\frac{K_{\mathcal{S}}}{2\eta(\varepsilon_{\mathcal{O}})}\right\}\\&>1-\exp\left(-(K-1)\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{(2\eta(\varepsilon_{\mathcal{O}})-1)nt^2}{2\eta(\varepsilon_{\mathcal{O}})K\sigma^2}\right)\end{aligned} \tag{18}$$

Further, due to Relation (14), we have $\frac{(2\eta(\varepsilon_{\mathcal{O}})-1)nt^2}{2\eta(\varepsilon_{\mathcal{O}})\sigma^2} < K \leq n$ and $\frac{K\sigma^2}{nt^2} \leq 1$; then, the inequality (18) can be bounded as

$$\begin{aligned}&\mathbb{P}\left\{\sum_{i\in[K_{\mathcal{S}}]}\mathbf{1}_{\{|\hat{\mu}_i-\mu_0|\leq t\}} \geq \frac{K_{\mathcal{S}}}{2\eta(\varepsilon_{\mathcal{O}})}\right\}\\ &> 1-\exp\left(-\left(\frac{(2\eta(\varepsilon_{\mathcal{O}})-1)nt^2}{2\eta(\varepsilon_{\mathcal{O}})\sigma^2}-1\right)\frac{2\eta(\varepsilon_{\mathcal{O}})-1}{2\eta(\varepsilon_{\mathcal{O}})}\log\frac{(2\eta(\varepsilon_{\mathcal{O}})-1)}{2\eta(\varepsilon_{\mathcal{O}})}\right)\end{aligned}$$

□

5. Conclusions

In this paper, we obtain the bounds of variance-dependen MoM estimation based on the binomial tail probability, including the case without pollution and the case with pollution. The nonasymptotic properties of nonpolluting MoM estimates have been shown to be superior to the existing traditional Hoeffding results. In the next step, we will also continue to investigate the bound of variance-dependen MoM estimation with outliers based on sub-Gaussian distribution or Weibull distribution. Compared with traditional exponential family distributions, it is more practical to study the inequalities of these distributions (see Refs. [19,20]). We further plan to study application problems with a practical background.

Author Contributions: Conceptualization, G.T. and Y.L.; methodology, G.T. and Y.L.; formal analysis, B.T.; writing—original draft preparation, G.T.; writing—review and editing, Y.L. and J.L.; supervision, B.T.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by China Postdoctoral Science Foundation 2023M733852.

Data Availability Statement: This paper does not use any data.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Nemirovskij, A.S.; Yudin, D.B. *Problem Complexity and Method Efficiency in Optimization*; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 1983.
2. Lerasle, M.; Szabó, Z.; Mathieu, T.; Lecué, G. Monk outlier-robust mean embedding estimation by median-of-means. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 3782–3793.
3. Lugosi, G.; Mendelson, S. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc.* **2019**, *22*, 925–965. [CrossRef]
4. Lecué, G.; Lerasle, M. Robust machine learning by median-of-means: Theory and practice. *Ann. Stat.* **2020**, *48*, 906–931. [CrossRef]
5. Humbert, P.; Le Bars, B.; Minvielle, L. Robust kernel density estimation with median-of-means principle. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MA, USA, 17–23 July 2022; p. 9444.
6. Wainwright, M.J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*; Cambridge University Press: Cambridge, UK, 2019; Volume 48.
7. Zhang, H.; Chen, S.X. Concentration Inequalities for Statistical Inference. *Commun. Math. Res.* **2021**, *37*, 1–85. [CrossRef]
8. Zhang, H.; Lei, X. Growing-dimensional Partially Functional Linear Models: Non-asymptotic Optimal Prediction Error. *Phys. Scr.* **2023**, *98*, 095216. [CrossRef]
9. Cai, T.T.; Han, R.; Zhang, A.R. On the non-asymptotic concentration of heteroskedastic Wishart-type matrix. *Electron. J. Probab.* **2022**, *27*, 1–40. [CrossRef]
10. Depersin, J.; Lecué, G. On the robustness to adversarial corruption and to heavy-tailed data of the Stahel–Donoho median of means. *Inf. Inference J. IMA* **2023**, *12*, 814–850. [CrossRef]
11. Marteau, C.; Sart, M. Deconvolution for some singular density errors via a combinatorial median of means approach. *Math. Stat. Learn.* **2023**, *6*, 51–85. [CrossRef]
12. Chen, Y. *A Short Note on the Median-of-Means Estimator*; University of Washington: Washington, DC, USA, 2020. Available online: https://faculty.washington.edu/yenchic/short_note/note_MoM.pdf (accessed on 12 November 2020).
13. Minsker, S. U-statistics of growing order and sub-Gaussian mean estimators with sharp constants. *arXiv* **2022**, arXiv:2202.11842.
14. Ferrante, G.C. Bounds on Binomial Tails With Applications. *IEEE Trans. Inf. Theory* **2021**, *67*, 8273–8279. [CrossRef]

15. Alsmeyer, G. Chebyshev's Inequality. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011. [CrossRef]
16. Lerasle, M. Lecture Notes: Selected Topics on Robust Statistical Learning Theory. *arXiv* **2019**, arXiv:1908.10761.
17. Zhang, H.; Wei, H.; Cheng, G. Tight Non-asymptotic Inference via Sub-Gaussian Intrinsic Moment Norm. *arXiv* **2023**, arXiv:2303.07287.
18. Depersin, J.; Lecué, G. Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Stat.* **2022**, *50*, 511–536. [CrossRef]
19. Hallinan, A.J., Jr. A review of the Weibull distribution. *J. Qual. Technol.* **1993**, *25*, 85–93. [CrossRef]
20. Xu, L.; Yao, F.; Yao, Q.; Zhang, H. Non-Asymptotic Guarantees for Robust Statistical Learning under Infinite Variance Assumption. *J. Mach. Learn. Res.* **2023**, *24*, 1–46.