A Robust Univariate Mean Estimator is All You Need

Adarsh Prasad[‡] **Sivaraman Balakrishnan**[†] **Pradeep Ravikumar**[‡]
Machine Learning Department[‡]
Department of Statistics and Data Science[†]
Carnegie Mellon University

Abstract

We study the problem of designing estimators when the data has heavy-tails and is corrupted by outliers. In such an adversarial setup, we aim to design statistically optimal estimators for flexible non-parametric distribution classes such as distributions with bounded-2k moments and symmetric distributions. Our primary workhorse is a conceptually simple reduction from multivariate estimation to univariate estimation. Using this reduction, we design estimators which are optimal in both heavy-tailed and contaminated settings. Our estimators achieve an optimal dimension independent bias in the contaminated setting, while also simultaneously achieving high-probability error guarantees with optimal sample complexity. These results provide some of the first such estimators for a broad range of problems including Mean Estimation, Sparse Mean Estimation, Covariance Estimation, Sparse Covariance Estimation and Sparse PCA.

1 Introduction

Modern data sets that arise in various branches of science and engineering are characterized by their ever increasing scale and richness. This has been spurred in part by easier access to computer, internet and various sensor-based technologies that enable the automated acquisition of such heterogeneous datasets. On the flip side, these large and rich data-sets are no longer carefully curated, are often collected in a decentralized, distributed fashion, and consequently are plagued with the complexities of heterogeneity, adversarial manipulations, and outliers. The analysis of these huge datasets is thus fraught with methodological challenges.

To understand the fundamental challenges and tradeoffs in handling such "dirty data" is precisely the premise of the field of robust statistics. Here, the aforementioned complexities are largely formalized under two different models of robustness: (1) **The heavy-tailed model:** In this model the sampling distribution can have thick tails, for instance, only low-order moments of the distribution are assumed to be finite; and (2) **The ϵ-contamination model:** Here the sampling distribution is modeled as a well-behaved distribution contaminated by an ϵ fraction of arbitrary outliers. In each case, classical estimators of the distribution (based for instance on the maximum likelihood estimator) can behave considerably worse (potentially arbitrarily worse) than under standard settings where the data is better behaved, satisfying various regularity properties. In particular, these classical estimators can be extremely sensitive to the tails of the distribution or to the outliers and the broad goal in robust statistics is to construct estimators that improve on these classical estimators by reducing their sensitivity to outliers.

Heavy Tailed Model. Concretely, focusing on the fundamental problem of robust mean estimation, in the heavy tailed model we observe n samples $x_1, \ldots, x_n$ drawn independently from a distribution P, which is only assumed to have low-order moments finite (for instance, P only has finite variance). The goal of past work Catoni (2012); Minsker (2015); Lugosi and Mendelson (2017); Catoni and Giulini (2017) has been to design an estimator $\widehat{\theta}_n$ of the true mean μ of P which has a small ℓ_2-error with high-probability. Formally, for a given $\delta > 0$, we would like an estimator with minimal r_δ such that,

$$P(\|\widehat{\theta}_n - \mu\|_2 \leq r_\delta) \geq 1 - \delta. \tag{1}$$

As a benchmark for estimators in the heavy-tailed model, we observe that when P is a multivariate normal distribution (or more generally is a sub-Gaussian

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

distribution) with mean μ and covariance Σ, it can be shown (see Hanson and Wright (1971)) that the sample mean $\widehat{\mu}_n = (1/n)\sum_i x_i$ satisfies, with probability at least $1-\delta$[1],

$$\|\widehat{\mu}_n - \mu\|_2 \lesssim \sqrt{\frac{\text{trace}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}}. \quad (2)$$

where $\|\Sigma\|_2$ denotes the operator norm of the covariance matrix Σ.

The bound is referred to as a *sub-Gaussian*-style error bound. However, for heavy tailed distributions, as for instance showed in Catoni (2012), the sample mean only satisfies the sub-optimal bound $r_\delta = \Omega(\sqrt{d/n\delta})$. Somewhat surprisingly, recent work Lugosi and Mendelson (2017) showed that the sub-Gaussian error bound is achievable while *only assuming that P has finite variance*, but by a carefully designed estimator. In the univariate setting, the classical median-of-means estimator Alon et al. (1996); Nemirovski and Yudin (1983); Jerrum et al. (1986) and Catoni's M-estimator Catoni (2012) achieve this surprising result but designing such estimators in the multivariate setting has proved challenging. Estimators that achieve truly sub-Gaussian bounds, but which are computationally intractable, were proposed recently by Lugosi and Mendelson (2017) and subsequently Catoni and Giulini (2017). Hopkins (2018) and Cherapanamjeri et al. (2019) developed a sum-of-squares based relaxation of Lugosi and Mendelson (2017)'s estimator, thereby giving a polynomial time algorithm which achieves optimal rates.

Huber's ϵ-Contamination Model. In this setting, instead of observing samples directly from the true distribution P, we observe samples drawn from P_ϵ, which for an arbitrary distribution Q is defined as a mixture model,

$$P_\epsilon = (1-\epsilon)P + \epsilon Q. \quad (3)$$

The distribution Q allows one to model arbitrary outliers, which may correspond to gross corruptions, or subtle deviations from the true model. There has been a lot of classical work studying estimators in the ϵ-contamination model under the umbrella of robust statistics (see for instance Hampel et al. (1986) and references therein). However, most of the estimators come that come with strong guarantees are computationally intractable Tukey (1975), while others are statistically sub-optimal heuristics Hastings et al. (1947). Recently, there has been substantial progress Diakonikolas et al. (2016); Lai et al. (2016); Kothari et al. (2018); Charikar et al. (2017); Diakonikolas et al. (2017); Balakrishnan et al. (2017); Prasad et al. (2018); Diakonikolas et al. (2018) designing provably robust which are computationally tractable while achieving near-optimal contamination dependence (i.e. dependence on the fraction of outliers ϵ) for computing means and covariances. In the Huber model, using information-theoretic lower bounds Chen et al. (2016); Lai et al. (2016); Hopkins and Li (2018), it can be shown that any estimator must suffer a *non-zero* bias (the asymptotic error as the number of samples go to infinity). For example, for the class of distributions with bounded variance, $\Sigma \precsim \sigma^2 \mathcal{I}_p$, the bias of any estimator is lower bounded by $\Omega(\sigma\sqrt{\epsilon})$. Surprisingly, the optimal bias that can be achieved is often independent of the data dimension. In other words, in many interesting cases optimally robust estimators in Huber's model can tolerate a constant fraction ϵ of outliers, *independent of the dimension.*

While the aforementioned recent estimators for mean estimation under Huber contamination have a polynomial computational complexity, their corresponding sample complexities are only known to be *polynomial* in the dimension p. For example, Kothari et al. (2018) and Hopkins and Li (2018) designed estimators which achieve optimal bias for distributions with *certifiably* bounded $2k$-moments, but their statistical sample complexity scales as $O(p^k)$. Steinhardt et al. (2017) studied mean estimation and presented an estimator which has a sample complexity of $\Omega\left(p^{1.5}\right)$.

Despite their apparent similarity, developments of estimators that are robust in each of these models, have remained relatively independent. Focusing on mean estimation we notice subtle differences, in the heavy-tailed model our target is the mean of the sampling distribution whereas in the Huber model our target is the mean of the *decontaminated* sampling distribution P. Beyond this distinction, it is also important to note that as highlighted above the natural focus in heavy-tailed mean estimation is on achieving strong, high-probability error guarantees, while in Huber's model the focus has been on achieving dimension independent bias.

Contributions. In this work, we aim to design estimators which are statistically optimally robust in both models simultaneously, *i.e.* they achieve a dimension-independent asymptotic bias in the ϵ-contamination model and achieve high probability deviation bounds similar to (2). Our main workhorse is a conceptually simple way of reducing multivariate estimation to the univariate setting. Then, by carefully solving mean estimation in the univariate setting, we are able to design optimal estimators for multivariate mean and covariance estimation for non-parametric distribution

[1]Here and throughout our paper we use the notation $\lesssim$ to denote an inequality with universal constants dropped for conciseness.

classes both in the low-dimensional ($n \geq p$) and high-dimensional ($n < p$) setting. We achieve these rates for non-parametric distribution classes such as distributions with bounded $2k$-moments and the class of symmetric distributions.

2 Background and Setup

In this section, we formally define two classes of distributions which we work with in this paper, (1) Bounded-$2k$-Moment distributions and (2) Symmetric Distributions.

Bounded $2k$-moment Class. Let x be a random vector with mean μ and covariance Σ. We say that x has bounded $2k$-moments if for all $v \in \mathcal{S}^{p-1}$, $\mathbb{E}[(v^T(x-\mu))^{2k}] \leq C_{2k}\left(\mathbb{E}[(v^T(x-\mu))^2]\right)^k$. We let $\mathcal{P}_{2k}^{\sigma^2}$ be the class of distributions with bounded $2k$ moments with covariance matrix $\Sigma \lesssim \sigma^2 \mathcal{I}_p$.

Symmetric Distributions. There exist several notions of symmetry for multivariate distributions. We discuss these notions briefly, but refer the reader to Liu (1990) for a detailed discussion. A random vector in $\mathbb{R}^p$ is centrally symmetric about $\theta \in \mathbb{R}^p$, if, $x - \theta \overset{\mathrm{d}}{=} \theta - x$, where $\overset{\mathrm{d}}{=}$ denotes *equal in distribution*. Equivalently, this corresponds to $u^T(x-\theta) \overset{\mathrm{d}}{=} u^T(\theta - x)$ for all unit vectors $u \in \mathcal{S}^{p-1}$. Liu (1990) introduced the broader notion of angular symmetry, where a random vector $x \in \mathbb{R}^p$ is angularly symmetric about θ, if $\frac{x-\theta}{\|x-\theta\|_2} \overset{\mathrm{d}}{=} \frac{\theta-x}{\|x-\theta\|_2}$, or equivalently, $\frac{x-\theta}{\|x-\theta\|_2}$ is centrally-symmetric. Central symmetry about θ implies angular symmetry about θ (see Lemma 2.2 in Liu (1990)).

Halfspace(H)-Symmetry. For any unit vector $u \in \mathcal{S}^{p-1}$, let $H_{u,t} = \{x : u^T x \leq t\}$ be a closed halfspace in $\mathbb{R}^p$. Its interior is an open subspace and the boundary $\{x : u^T x = t\}$ is a hyperplane. Recall that for any random variable $y \in \mathbb{R}$, the *median* of the distribution of y ($\mathrm{med}(y)$) is defined to be any number c such that $\Pr(y \leq c) \geq 0.5, \quad \Pr(y \geq c) \geq 0.5$. Then, a random vector in $\mathbb{R}^p$ is H-symmetric about $\theta \in \mathbb{R}^p$ if, $\Pr(x \in H) \geq \frac{1}{2}$ for all closed halfspaces H with θ on boundary. Note that angular symmetry about a point θ implies halfspace-symmetry about it as well (see Lemma 2.4 Liu (1990)). Moreover, if we have that x is H-symmetric about θ, then (1) $\mathrm{med}(u^T x) = u^T\theta$, and (2) $\Pr(u^T(x-\theta) \geq 0) \geq \frac{1}{2}$ for all $u \in \mathcal{S}^{p-1}$ (see Theorem 2.1 Liu (1990)). Note that till now, our discussion hasn't required the distribution to have bounded moments, in particular, it need not even have bounded first moments (mean). However, if the distribution has a finite mean, then, it is easy to see that $\mathrm{med}(u^T x) = \mathbb{E}[u^T x] = u^T\theta$. Our last assumption ensures that the median is unique and hence identifiable. To this end, let $\mathcal{P}_{\mathrm{sym}}$ be the class of *H-symmetric* distributions with unique center of H-symmetry. Moreover suppose $\mathcal{P}_{\mathrm{sym}}^{t_0,\kappa} \subset \mathcal{P}_{\mathrm{sym}}$ is the class of distributions such that for any $P \in \mathcal{P}_{\mathrm{sym}}^{t_0,\kappa}$ the CDF of the univariate projection$(u^T P)$ given by $F_{u^T P}$ increases at least linearly around $u^T\theta$. Formally, for all $x_1 \in [\mathrm{med}(u^T P), F_{u^T P}^{-1}(\frac{1}{2}+t_0)]$ we have that

$$F_{u^T P}(x_1) - \frac{1}{2} \geq \frac{1}{\kappa_{u,P}}(x_1 - \mathrm{med}(u^T P)) \tag{4}$$

and for all $x_2 \in [F_{u^T P}^{-1}(\frac{1}{2}-t_0), \mathrm{med}(u^T P)]$, we have that $\frac{1}{2} - F_{u^T P}(x_2) \geq \frac{1}{\kappa_{u,P}}(\mathrm{med}(u^T P) - x_2)$ for $\kappa_{u,P} \leq \kappa$. A higher κ corresponds to slower rate of increase in CDF around the median. Note that κ can be thought of as a measure of variance or dispersion. In particular, for example, in the case of univariate Gaussian distribution, *i.e.* $P = \mathcal{N}(\mu, \sigma^2)$, $\kappa(P) = C\sigma$. Similarly for univariate Cauchy distribution with scale γ, $\kappa(P) \approx C\gamma$. Note that any univariate distribution $P \in \mathcal{P}_{\mathrm{sym}}$ with density function $p(x)$ such that $\min_{|t|<t_0} p(P_F^{-1}(\frac{1}{2}+t)) > \frac{1}{\kappa}$ also belongs to $\mathcal{P}_{\mathrm{sym}}^{t_0,\kappa}$.

3 Some Candidate Multivariate Estimators

In this section, we study some natural candidate estimators, to see if they achieve an optimal asymptotic bias in the ϵ-contamination model. We assume that the true distribution is a multivariate isotropic gaussian, $P = \mathcal{N}(0, \mathcal{I}_p)$. Observe that it lies in both $\mathcal{P}_{2k}^{\sigma^2}$ and $\mathcal{P}_{\mathrm{sym}}^{t_0,\kappa}$ for $\sigma^2 = 1$, and $\kappa = O(1)$, hence our results for both distribution classes.

Convex M-estimation. M-estimators were originally proposed by Huber Huber (1965), and were shown to be robust in one dimension. Subsequent research in 1970s showed that M-estimators perform poorly for multivariate data Maronna (1976). In particular, Donoho and Gasko (1992) showed that when the data is p-dimensional, the breakdown point of M-estimators scales inversely with the dimension. Lai et al. (2016) and Diakonikolas et al. (2016) derived similar negative results for the specific case of geometric median. We further extend this observation, and show that even at a very small contamination level, *i.e.* $\epsilon \mapsto 0$, the bias of certain convex M-estimators which are Fisher-consistent for $\mathcal{N}(0, \mathcal{I}_p)$ will necessarily scale polynomially in the dimension.

Lemma 1. *Let $P = \mathcal{N}(0, \mathcal{I}_p)$ and consider the convex risk $R_P(\theta) = \mathbb{E}_{z\sim P}[\ell(\|z-\theta\|_2)]$ where $\ell : \mathbb{R} \mapsto \mathbb{R}$ be any twice differentiable Fisher-consistent convex loss,* i.e. *$\theta(P) = \mathrm{argmin}_\theta R_P(\theta) = 0$. Then, there exists a corruption Q such that $\lim_{\epsilon \mapsto 0} \|\theta(P_\epsilon)\|_2 \geq \epsilon\sqrt{p}$.*

Recall that when the true distribution $P = \mathcal{N}(0, \mathcal{I}_p)$, then, the lower bound on estimation in the Huber

Model is $\Theta(\epsilon)$ Chen et al. (2015). Our explicit dimension-dependent lower bound on the bias of M-estimators shows their sub-optimality.

Subset Search. Having ruled out convex estimation to a certain extent, we next turn our attention to non-convex methods. Perhaps the most simple non-convex method is simple search. Intuitively, the squared loss measures the *fit* between a parameter θ and samples $\mathcal{Z}$, and if all samples don't come from the same distribution(*i.e.* have outliers), then the corresponding *fit* should be bad. To capture this intuition algorithmically, one can (1) consider all subsets of size $\lfloor(1-\epsilon)n\rfloor$, (2) minimize the squared loss over these subsets, and then (3) return the estimator corresponding to the subset with least squared loss or best fit. To be precise, given n samples from P_ϵ

$$S^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{S \ s.t. \ |S|=(1-\epsilon)n} \min_\theta \frac{1}{(1-\epsilon)n} \sum_{x_i \in S} \|x_i - \theta\|_2^2$$

$$\widehat{\theta}_{\text{SRM}} \stackrel{\text{def}}{=} \min_\theta \frac{1}{(1-\epsilon)n} \sum_{x_i \in S^*} \|x_i - \theta\|_2^2 \tag{5}$$

Our next result studies the asymptotic performance of this estimator.

Lemma 2. *Let $P = \mathcal{N}(0, \mathcal{I}_p)$, then as $n \mapsto \infty$, we have that*

$$\sup_Q \|\widehat{\theta}_{SRM} - \mathbb{E}_{x\sim P}[x]\|_2 = \frac{\epsilon}{\sqrt{(1-\epsilon)(1-2\epsilon)}}\sqrt{p}. \tag{6}$$

The above result shows that, while subset-search has a finite dimension-independent breakdown point(0.5), the bias of this estimator necessarily scales with the dimension $\sqrt{p}$.

4 Optimal Univariate Estimation

In the previous section, we studied some natural candidate estimators and showed that they don't achieve the optimal asymptotic bias in ϵ-contamination model for multivariate mean estimation. In this section, we take a step back, and study univariate estimation.

4.1 Bounded 2k-moments

We study the interval estimator which was initially proposed by Lai et al. (2016). The estimator, presented in Algorithm 1, proceeds by using half of the samples to identify the shortest interval containing at least $(1-\epsilon)n$ fraction of the points, and then the remaining half of the points is used to return an estimate of the mean.

We assume that the contamination level ϵ and confidence level δ are such that,

$$2\epsilon + \sqrt{\epsilon\frac{\log(4/\delta)}{n}} + \frac{\log(4/\delta)}{n} < \frac{1}{2}.$$

Algorithm 1 Robust Univariate Mean Estimation

function INTERVAL1D($\{z_i\}_{i=1}^{2n}$, CORRUPTION LEVEL ϵ, CONFIDENCE LEVEL δ)

Split the data into two subsets: $\mathcal{Z}_1 = \{z_i\}_{i=1}^n$ and $\mathcal{Z}_2 = \{z_i\}_{i=n+1}^{2n}$.

Let $\alpha = \max(\epsilon, \frac{\log(1/\delta)}{n})$.

Using $\mathcal{Z}_1$, let $\hat{I} = [a, b]$ be the shortest interval containing $n(1-2\alpha-\sqrt{2\alpha\frac{\log(4/\delta)}{n}} - \frac{\log(4/\delta)}{n})$ points.

Use $\mathcal{Z}_2$ to identify points lying in $[a, b]$.

return $\frac{1}{\sum_{i=n}^{2n}\mathbb{I}\left\{z_i\in\hat{I}\right\}}\sum_{i=n}^{2n} z_i\mathbb{I}\left\{z_i \in \hat{I}\right\}$

end function

Then, we have the following Lemma.

Lemma 3. *Suppose P be any $2k$-moment bounded distribution over $\mathbb{R}$ with mean μ with variance bounded by σ^2. Given, n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (3), Algorithm 1 returns an estimate $\widehat{\theta}_\delta$ such that with probability at least $1-\delta$,*

$$|\widehat{\theta}_\delta - \mu| \lesssim \sigma \max(2\epsilon, \frac{\log(1/\delta)}{n})^{1-\frac{1}{2k}} + \sigma(\frac{\log n}{n})^{1-\frac{1}{2k}} + \sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

Observe that Algorithm 1 has an asymptotic bias of $O(\sigma\epsilon^{1-1/2k})$ in the ϵ-contamination setting, which is known to be information theoretically optimal Hopkins and Li (2018); Lai et al. (2016).

Moreover, observe that for $\epsilon = 0$, P has atleast bounded 4th moment, *i.e.* $k \geq 2$, $\frac{\log(n)}{n}^{1-1/2k}$ term can be ignored for large enough n. Hence, for $k \geq 2$ and large enough n, Algorithm 1 achieves the deviation rate of $\sigma\sqrt{\frac{\log(1/\delta)}{n}}$.

4.2 Symmetric Distributions

In the univariate setting, our estimator presented in Algorithm 2 simply returns the sample median of the observed samples. While this idea is simple and crucially exploits that the mean and median overlap for a symmetric distribution, this leads to profound implications on the effect of contamination in the Huber contamination model. Next, we present the theoretical bound achieved by this estimator, which was shown in Altschuler et al. (2018).

We further assume that ϵ and δ are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{\log(2/\delta)}{n}} \leq t_0.$$

Then we have that,

Algorithm 2 Sample Median

function SAMPLE MEDIAN - 1D ($\{z_i\}_{i=1}^{2n+1}$)
  Let $z_{[k]}$ be k^{th} order-statistic
  return $z_{[n+1]}$
end function

Lemma 4. *[Altschuler et al. (2018)] Let P be any univariate distribution in $\mathcal{P}_{sym}^{t_0,\kappa}$. Given n-samples from the mixture distribution (3), we get that with probability at least $1-\delta$, Algorithm 2 returns an estimate $\widehat{\theta}$ such that,*

$$|\widehat{\theta} - \mathbb{E}_{x\sim P}[x]| \leq C_1\kappa\epsilon + C_2\kappa\sqrt{\frac{\log(1/\delta)}{n}}$$

Observe that Algorithm 2 has an asymptotic bias of $O(\kappa\epsilon)$, which is also information theoretically optimal. To see this, observe that $\mathcal{N}(\cdot,\kappa^2)$ lies in $\mathcal{P}_{\text{sym}}^{t_0,\kappa}$ for some constant t_0 and the fact that $TV(\mathcal{N}(\mu_1,\kappa^2),\mathcal{N}(\mu_2,\kappa^2)) = O(|\mu_1-\mu_2|/\kappa)$(Theorem 1.3 Devroye et al. (2018)).

5 From 1D to p-D: A meta-estimator

In this section, we propose a general meta-estimator to extend any univariate estimator to the multivariate setting. For any univariate estimator $f(\cdot)$, suppose that when given n-samples from the mixture model, it returns an estimate $f(\mathcal{X}_n)$ such that with probability $1-\delta$,

$$|f(\mathcal{X}_n,\epsilon,\delta) - \mu(P)| \leq \omega_f(\epsilon,P,\delta).$$

Note that $\omega_f(\epsilon,P,\delta)$ is the error suffered by the univariate estimator at a contamination level ϵ, and confidence level δ, when the true univariate distribution is P.

5.1 Mean Estimation

The proposed meta-estimator proceeds by robustly estimating the mean along almost every direction u, and returns an estimate $\widehat{\theta}$, whose projection along $u(u^T\widehat{\theta})$ is close to these univariate robust mean along that direction. In particular, let $\mathcal{N}^{1/2}(\mathcal{S}^{p-1})$ is the half-cover of the unit sphere $\mathcal{S}^{p-1}$, *i.e.* $\forall u \in \mathcal{S}^{p-1}$, there exists a $y \in \mathcal{N}^{1/2}(\mathcal{S}^{p-1})$ such that $u = y+z$ for some $\|z\|_2 \leq \frac{1}{2}$. Then, for any point $\theta \in \mathbb{R}^p$ and *any* univariate estimator $f(\cdot)$, consider the following loss,

$$D_{\text{f}}(\theta,\{x_i\}_{i=1}^n) = \sup_{u\in\mathcal{N}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \text{f}(\{u^Tx_i\}_{i=1}^n,\epsilon,\frac{\delta}{5^p})|,$$

Then, we use it to construct the following multivariate meta-estimator, $\widehat{\theta}_f$ which takes in n-samples $\{x_i\}_{i=1}^n$ and a univariate estimator $f(\cdot)$,

$$\widehat{\theta}_{\text{f}}(\{x_i\}_{i=1}^n) = \operatorname*{argmin}_{\theta} D_f(\theta,\{x_i\}_{i=1}^n),$$

Such directional-control based estimators have been previously studied in the context of heavy-tailed mean estimation by Joly et al. (2017) and Catoni and Giulini (2017). Joly et al. (2017) used the median-of-means estimator, while Catoni and Giulini (2017) used Catoni's M-estimator Catoni (2012) as their univariate estimator. Then, we have the following result.

Lemma 5. *Suppose P is a multivariate distribution with mean μ. Given n-samples from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_f(\mathcal{X}_n) - \mu\|_2 \lesssim \sup_{u\in\mathcal{N}^{1/2}(\mathcal{S}^{p-1})} \omega_f(\epsilon,u^TP,\frac{\delta}{5^p}),$$

where u^TP is the univariate distribution of P along u.

Sparse Mean Estimation. In this setting, we further assume that the true mean vector of the distribution P has only a few non-zero co-ordinates, *i.e.* it is sparse. Such sparsity patterns are known to be present in high-dimensional data(see Rish et al. (2014) and references therein). Then, the goal is to design estimators which can exploit this sparsity structure, while remaining robust under the ϵ-contamination model. Formally, for a vector $x \in \mathbb{R}^p$, let $\text{supp}(x) = \{i \in [p] \;\; s.t. \;\; x(i) \neq 0\}$. Then, x is s-sparse if $|\text{supp}(x)| \leq s$. We further assume that $s \leq p/2$. Let Θ_s be the set of s-sparse vectors in $\mathbb{R}^p$, and let $\mathcal{N}_{2s}^{\frac{1}{2}}(\mathcal{S}^{p-1})$ is the half-cover of the set of unit vectors which are 2s-sparse. Then, for any univariate estimator $f(\cdot)$, let

$$D_{f,s}(\theta,\{x_i\}_{i=1}^n) = \sup_{u\in\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \text{f}(u^T\mathcal{X}_n,\epsilon,\frac{\delta}{(\frac{6ep}{s})^s})|.$$

Then, we can define the following meta-estimator,

$$\widehat{\theta}_{\text{f,s}}(\mathcal{X}_n) = \operatorname*{argmin}_{\theta\in\Theta_s} D_{f,s}(\theta,\mathcal{X}_n),$$

which has the following error guarantee.

Lemma 6. *Suppose P is a multivariate distribution with mean μ such that μ is s-sparse. Given n-samples from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{f,s}(\mathcal{X}_n) - \mu\|_2 \lesssim \sup_{u\in\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} \omega_f(\epsilon,u^TP,\frac{\delta}{(\frac{6ep}{s})^s}),$$

where u^TP is the univariate distribution of P along u.

5.2 Covariance-Estimation

In this section, we study recovering the true covariance matrix, when given samples from a mixture distribution. We first center our observations by defining pseudo-samples $z_i = \frac{x_i - x_{i+n/2}}{\sqrt{2}}$. We can think of z_i

as being sampled from the Huber Contamination $\tilde{P}_{2\epsilon}$, where $\tilde{P} = \frac{1}{\sqrt{2}}(P - P)$. Let $\mathcal{Z}_n = \{z_i\}_{i=1}^n$ be the set of these pseudo-samples, and let $u^T\mathcal{Z}_n^{\otimes 2} = \{(u^T z_i)^2\}_{i=1}^n$. Let $\mathcal{F} = \{\Sigma = \Sigma^T \in \mathbb{R}^{p\times p} : \Sigma \succeq 0\}$ be the class of positive semi-definite symmetric matrices. Then, for any matrix $\Theta \in \mathcal{F}$, let,

$$\mathcal{D}_{\mathrm{f}}^{\otimes 2}(\Theta, \mathcal{Z}_n) = \sup_{u\in\mathcal{N}^{1/4}} |u^T\Theta u - \mathrm{f}(u^T\mathcal{Z}_n^{\otimes 2}, \epsilon, \frac{\delta}{9^p})|$$

Then, consider the meta-estimator $\widehat{\Theta}_{\mathrm{f}}(\mathcal{Z}_n)$ given as,

$$\widehat{\Theta}_{\mathrm{f}}(\mathcal{Z}_n) = \operatorname*{argmin}_{\Theta\in\mathcal{F}} \mathcal{D}_{\mathrm{f}}^{\otimes 2}(\Theta, \mathcal{Z}_n)$$

Lemma 7. *Suppose P is a multivariate distribution with covariance Σ. Given n-samples from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\Theta}_f(\mathcal{Z}_n) - \Sigma\|_2 \lesssim \sup_{u\in\mathcal{N}^{1/4}(\mathcal{S}^{p-1})} \omega_f(2\epsilon, u^T\tilde{P}^{\otimes 2}, \frac{\delta}{9^p}),$$

where $u^T\tilde{P}^{\otimes 2}$ is the univariate distribution of $(u^Tz_i)^2$ for $z_i \sim \tilde{P}$.

Sparse Covariance Estimation. Next, we consider sparse covariance matrices. In particular, we assume that there exists a subset S of $|S| = s$ covariates that are correlated with each other, and the remaining covariates $[p]\backslash S$ are independent from this subset and from each other. Such sparsity patterns arise naturally in various real-world data Bien and Tibshirani (2011). More concretely, for a subset of co-ordinates S, define $\mathcal{G}(S) \stackrel{\text{def}}{=} \{G = (g)_{ij} \in \mathbb{R}^{p\times p}, g_{ij} = 0 \text{ if } i \notin S \text{ or } j \notin S\}$, and let $\mathcal{G}(s) = \bigcup_{S\subset[p]:|S|\le s} \mathcal{G}(S)$. Consider the class of matrices $\mathcal{F}_s$ such that,

$$\mathcal{F}_s = \{\Sigma = \Sigma^T, \Sigma \succeq 0, \Sigma - \mathrm{diag}\Sigma \in \mathcal{G}(s)\}$$

Then for any matrix Θ and univariate estimator f, let

$$D_{\mathrm{f},s}(\Theta, \mathcal{Z}_n) = \sup_{u\in\mathcal{N}_{2s}^{1/4}(\mathcal{S}^{p-1})} |u^T\Theta u - \mathrm{f}(u^T\mathcal{Z}_n^{\otimes 2}, \epsilon, \frac{\delta}{(\frac{9ep}{s})^s})|.$$

Then, we can define the following estimator,

$$\widehat{\Theta}_{\mathrm{f,s}}(\mathcal{X}_n) = \operatorname*{arginf}_{\theta\in\mathcal{F}_s} D_{\mathrm{f},s}(\Theta, \mathcal{Z}_n),$$

Lemma 8. *Suppose P is a multivariate distribution with covariance Σ such that $\Sigma \in \mathcal{F}_s$. Given n-samples from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\Theta}_{f,s}(\mathcal{Z}_n) - \Sigma\|_2 \lesssim \sup_{u\in\mathcal{N}_{2s}^{1/4}(\mathcal{S}^{p-1})} \omega_f(2\epsilon, u^T\tilde{P}^{\otimes 2}, \frac{\delta}{(\frac{9ep}{s})^s}),$$

where $u^T\tilde{P}^{\otimes 2}$ is the univariate distribution of $(u^Tz_i)^2$ for $z_i \sim \tilde{P}$.

6 Consequences for $\mathcal{P}_{2k}^{\sigma^2}$

Next, we study the performance of our meta-estimator for multivariate estimation for the class of distributions with bounded $2k$-moments. In particular, we use the interval estimator(IM) presented in Algorithm 1 as our univariate estimator to instantiate our meta-estimator.

Multivariate Mean Estimation. In the multivariate setting, we further assume that the contamination level ϵ, and confidence are such that,

$$2\epsilon + \sqrt{\epsilon(\frac{p}{n} + \frac{\log(1/\delta)}{n})} + \frac{p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 1. *Suppose P has bounded $2k$ moments with mean μ and covariance Σ. Given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{IM}(\mathcal{X}_n) - \mu\|_2 \lesssim \|\Sigma\|_2^{1/2}\epsilon^{1-1/2k} + \|\Sigma\|_2^{1/2}\sqrt{\frac{\log(1/\delta)}{n}} + \|\Sigma\|_2^{1/2}(\sqrt{\frac{p}{n}} + (\frac{\log n}{n})^{1-\frac{1}{2k}})$$

Observe that the proposed estimator achieves a dimension independent asymptotic bias of $O(\sigma\epsilon^{1-1/2k})$ in the ϵ-contamination model for multivariate mean estimation, with a sample complexity of $O(p)$.

Sparse Mean Estimation. In this setting, we assume that the contamination level ϵ, and confidence are such that,

$$2\epsilon + \sqrt{\epsilon(\frac{s\log p}{n} + \frac{\log(1/\delta)}{n})} + \frac{s\log p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 2. *Suppose P has bounded $2k$ moments with mean μ and covariance Σ, where μ is s-sparse. Then, given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{IM,s}(\mathcal{X}_n) - \mu\|_2 \lesssim \|\Sigma\|_{2,2s}^{1/2}\epsilon^{1-1/2k} + \|\Sigma\|_{2,2s}^{1/2}\sqrt{\frac{\log(1/\delta)}{n}} + \|\Sigma\|_{2,2s}^{1/2}(\sqrt{\frac{s\log p}{n}} + (\frac{\log n}{n})^{1-\frac{1}{2k}}),$$

where $\|\Sigma\|_{2,2s} = \sup_{u\in\mathcal{S}^{p-1},\|u\|_0\le 2s} u^T\Sigma u$.

The above result shows that the proposed estimator exploits the underlying sparsity structure, and achieves the near-optimal sample complexity of

$O(s \log p)$, while simultaneously achieving the optimal asymptotic bias of $O(\|\Sigma\|_{2,2s}^{1/2}\epsilon^{1-1/2k})$.

Covariance Estimation. We begin by first calculating $\omega_{IM}(2\epsilon, u^T\widetilde{P}^{\otimes 2}, \delta)$. To do this, recall that for fixed u, for the clean samples in z_i, $(u^T z_i)^2$ has mean $u^T\Sigma(P)u$, and variance $C_4(u^T\Sigma(P)u)^2$. Note that the scalar random variables $(u^T z_i)^2$ have bounded k moments, whenever x_i has bounded $2k$-moments. Hence, from Lemma 3, we have that

$$\omega_{IM}(2\epsilon, u^T\widetilde{P}^{\otimes 2}, \delta) \lesssim (u^T\Sigma(P)u)\epsilon^{1-1/k} + u^T\Sigma(P)u\sqrt{\frac{\log 1/\delta}{n}}.$$

We assume that the contamination level ϵ, and confidence are such that,

$$4\epsilon + \sqrt{2\epsilon(\frac{p}{n} + \frac{\log(1/\delta)}{n})} + \frac{p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 3. *Suppose P has bounded $2k$-moments, then, given $\mathcal{X}_n$ drawn from the mixture model, then, we have that with probability at least $1-\delta$,*

$$\|\widehat{\Theta}_{IM} - \Sigma(P)\|_2 \lesssim \|\Sigma(P)\|_2\epsilon^{1-1/k} + \|\Sigma(P)\|_2\sqrt{\frac{p}{n}} + \|\Sigma(P)\|_2\sqrt{\frac{\log 1/\delta}{n}}$$

Observe that the proposed estimator achieves a dimension independent asymptotic bias of $O(\sigma^2\epsilon^{1-1/k})$ in the ϵ-contamination model for multivariate covariance estimation, with a sample complexity of $O(p)$.

Sparse Covariance Estimation. In this setting, we assume that the contamination level ϵ, and confidence δ are such that,

$$4\epsilon + \sqrt{2\epsilon(\frac{s\log p}{n} + \frac{\log(1/\delta)}{n})} + \frac{s\log p}{n} + \frac{\log(4/\delta)}{n} < c,$$

for some small constant $c > 0$. Then, we have the following result.

Corollary 4. *Suppose P has bounded $2k$-moments and $\Sigma(P) \in \mathcal{F}_s$, then, given $\mathcal{X}_n$ drawn from the mixture model, we have that with probability at least $1-\delta$,*

$$\|\widehat{\Theta}_{IM,s} - \Sigma(P)\|_2 \lesssim \|\Sigma(P)\|_2\epsilon^{1-1/k} + \|\Sigma(P)\|_2\sqrt{\frac{s\log p}{n}} + \|\Sigma(P)\|_2\sqrt{\frac{\log 1/\delta}{n}}$$

As before, even in this case, the proposed estimator achieves a dimension independent bias of $O(\sigma^2\epsilon^{1-1/k})$, with a sample complexity of $O(s\log p)$.

Sparse PCA in Spiked Covariance Model As an application of the sparse-covariance estimation, we consider the following spiked covariance model, where the true distribution $P \in \mathcal{P}_{2k}$ is such that

$$\Sigma(P) = V\Lambda V^T + \mathcal{I}_p, \tag{7}$$

where $V \in \mathbb{R}^{p\times r}$ is an orthonormal matrix, and $\Lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix with entries $\Lambda_1 \geq \Lambda_2 \geq \ldots \geq \Lambda_r > 0$. In this setting, suppose we observe samples from a mixture distribution P_ϵ, then the goal is to estimate the subspace projection matrix VV^T, *i.e.* construct $\widehat{V}$ such that $\|\widehat{V}\widehat{V} - VV^T\|_F$ is small. Note that when V has only s non-zero rows, then the corresponding covariance matrix Σ is s-sparse($\Sigma \in \mathcal{F}_s$).

We follow Chen et al. (2015) to use our sparse covariance estimator $\widehat{\Theta}_{\mathrm{IM,s}}(\mathcal{X}_n)$ to construct $\widehat{V} \in \mathbb{R}^{p\times r}$ by setting its j^{th} column to be the j^{th} eigenvector of $\widehat{\Theta}_{\mathrm{IM,s}}(\mathcal{X}_n)$. Then, under the assumption that (ϵ, n) are such that

$$(1+\Lambda_1)\epsilon^{1-1/k} + (1+\Lambda_1)\sqrt{\frac{s\log p}{n}} + (1+\Lambda_1)\sqrt{\frac{\log 1/\delta}{n}} \lesssim \frac{\Lambda_r}{2},$$

we have the following result.

Corollary 5. *Suppose P has bounded $2k$-moments, and $\Sigma(P)$ is of the form of (7) and we are given n samples from the mixture distribution. Then, we have that with probability at least $1-\delta$,*

$$\|\widehat{V}\widehat{V}^T - VV^T\|_F^2 \lesssim (\frac{1+\Lambda_1}{\Lambda_r})^2(\epsilon^{2-2/k}) + (\frac{1+\Lambda_1}{\Lambda_r})^2(\frac{s\log p}{n} + \frac{\log 1/\delta}{n})$$

Discussion. Throughout this section, all our estimators achieve a dimension-independent asymptotic bias. Hence, our proposed meta-estimator allows us to escape the dimension dependence in the ϵ-contamination setting.

Next, we expand on a more subtle aspect of our estimators. Observe that when $\epsilon = 0$, *i.e.* there is no contamination, we see that the typical error rate of our estimators for $k \geq 2$($k \geq 4$ for covariance) is $O(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}})$ is the low dimensional setting, and $O(\sqrt{\frac{s\log p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}})$ in the high-dimensional setting. Typically, such high-probability bounds are achieved only under the restrictive assumption that the true distribution is Gaussian or sub-gaussian. In contrast, all our results are valid for the much broader class of distributions with bounded $2k$-moment. As discussed in the introduction, while such results have been recently obtained for mean estimation, our simple meta-estimator achieves these high-probability error

guarantees for a much broader range of problems. To the best of our knowledge, these are some of the first estimators which get such high-probability deviation bounds for sparse-mean, covariance, sparse-covariance and sparse-PCA.

7 Consequences for $\mathcal{P}_{\mathbf{sym}}^{t_0,\kappa}$

Next, we study the performance of our meta-estimator for multivariate estimation for the class of symmetric distributions. In particular, we use the sample median presented in Algorithm 2 as our univariate estimator to instantiate our meta-estimator. In the multivariate setting, we further assume that the contamination level ϵ, and confidence level δ are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{p}{n} + \frac{\log(2/\delta)}{n}} \leq t_0.$$

Then, we have the following result.

Corollary 6. *Suppose $P \in \mathcal{P}_{sym}^{t_0,\kappa}$ is a multivariate distribution with mean μ. Given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{Med}(\mathcal{X}_n) - \mu\|_2 \lesssim \kappa\epsilon + \kappa\sqrt{\frac{\log(1/\delta)}{n}} + \kappa\sqrt{\frac{p}{n}}$$

Observe that the proposed estimator achieves a dimension independent asymptotic bias of $O(\kappa\epsilon)$ in the ϵ-contamination model for multivariate mean estimation, with a sample complexity of $O(p)$.

Sparse Mean Estimation. In this setting, we assume that the contamination level ϵ, and confidence are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{s\log p}{n} + \frac{\log(2/\delta)}{n}} \lesssim t_0.$$

Then, we have the following result.

Corollary 7. *Suppose $P \in \mathcal{P}_{sym}^{t_0,\kappa}$ is a multivariate distribution with mean μ. Given n samples $\{x_i\}_{i=1}^n$ from the mixture distribution (3), we get that with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{Med,s}(\mathcal{X}_n) - \mu\|_2 \lesssim \kappa\epsilon + \kappa\sqrt{\frac{\log(1/\delta)}{n}} + \kappa\sqrt{\frac{s\log p}{n}}$$

The above result shows that the proposed estimator exploits the underlying sparsity structure, and achieves the near-optimal sample complexity of $O(s\log p)$, while simultaneously achieving the optimal asymptotic bias of $O(\kappa\epsilon)$. Moreover for the case of $\epsilon = 0$, the proposed estimator achieves the near-optimal deviation bound for sparse-mean estimation, for symmetric distributions without moments. Note that similar results can be derived for other higher-order moments.

Discussion. Observe the difference in achievable rates for $\mathcal{P}_{\text{sym}}^{t_0,\kappa}$ and $\mathcal{P}_{2k}^{\sigma^2}$. In particular, for symmetric distributions including those which have no finite variance, the maximum bias introduced by Huber Contamination Model is at most $O(\kappa\epsilon)$. In contrast for distributions with bounded $2k$-moments, the lower bound for mean estimation is $\Omega(\sigma\epsilon^{1-1/2k})$. Note that the depth based estimators of Chen et al. (2015) also implicitly assume that the underlying distribution is symmetric, and hence obtain similar rates for elliptical distributions.

8 Conclusion and Future Directions.

In this work we provided a conceptually simple way of reducing multivariate estimation to univariate estimation. In particular, we showed how to use any robust univariate estimator to design statistically optimal robust estimators for multivariate estimation. Through this reduction, we derived optimal estimators for non-parametric distribution classes such as distributions with bounded $2k$-moments and symmetrical distributions. Our estimators achieved optimal asymptotic bias in the ϵ-contamination model, and also high-probability deviation bounds in the uncontaminated setting. There are several avenues for future work, some of which we discuss below.

Computationally Efficient Estimators. As noted in the introduction, there has been a flurry of work in the theoretical computer science community on designing polynomial time estimators for robust mean estimation. Designing sample-efficient estimators for sparse-mean estimation for the bounded $2k$-moment class is an open problem. Similarly for covariance estimation, most existing work has focused on Frobenius norm, or Mahalanobis distance, and designing estimators for covariance estimation in operator norm for general bounded $2k$-moment is an open problem. Another important challenge is to design computationally efficient estimators for the mean of a symmetric distribution.

Extension to General Risk Minimization. Another future line of work is to extend our results for general risk minimization problems such as Linear Regression and Generalized Linear Models.

9 Acknowledgements.

AP and PR acknowledge the support of NSF via IIS-1909816.

References

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*, 2017.

Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3): 1079–1083, 1971.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 20–29, New York, NY, USA, 1996. ACM.

A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43: 169 – 188, 1986.

Samuel B Hopkins. Sub-Gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-Gaussian rates. *arXiv preprint arXiv:1902.01998*, 2019.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics*. Wiley Online Library, 1986.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

Cecil Hastings, Frederick Mosteller, John W Tukey, and Charles P Winsor. Low moments for small samples: a comparative study of order statistics. *The Annals of Mathematical Statistics*, 18(3):413–426, 1947.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008, 2017.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for Huber's epsilon-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM, 2018.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.

Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.

Ricardo Antonio Maronna. Robust M-estimators of multivariate location and scatter. *The annals of statistics*, pages 51–67, 1976.

David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4): 1803–1827, 1992.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *ArXiv e-prints, to appear in the Annals of Statistics*, 2015.

Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *arXiv preprint arXiv:1802.09514*, 2018.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Emilien Joly, Gábor Lugosi, and Roberto Imbuzeiro Oliveira. On the estimation of the mean of a random vector. *Electronic Journal of Statistics*, 11(1):440–451, 2017.

Irina Rish, Guillermo A Cecchi, Aurelie Lozano, and Alexandru Niculescu-Mizil. *Practical applications of sparse modeling*. MIT Press, 2014.

Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Roman Vershynin. On the role of sparsity in compressed sensing and random matrix theory. In *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 189–192. IEEE, 2009.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.