Robust Mean Estimation on Highly Incomplete Data with Arbitrary Outliers

Lunjia Hu
Stanford University

Omer Reingold
Stanford University

Abstract

We study the problem of robustly estimating the mean of a d-dimensional distribution given N examples, where most coordinates of every example may be missing and εN examples may be arbitrarily corrupted. Assuming each coordinate appears in a constant factor more than εN examples, we show algorithms that estimate the mean of the distribution with information-theoretically optimal dimension-independent error guarantees in nearly-linear time $\widetilde{O}(Nd)$. Our results extend recent work on computationally-efficient robust estimation to a more widely applicable incomplete-data setting.

1 INTRODUCTION

Imagine a program committee that wants to evaluate a collection of papers with a group of reviewers. As it is often impractical for a single reviewer to read all the papers, the common practice is to assign every paper only to a small sub-group of the reviewers. Assuming that all the reviewers are reliable, and the ratings they provide for a paper p are independently drawn from a distribution $\mathcal{D}_p$, one can estimate the mean of the distribution simply by taking the average rating over the sub-group. The estimation error can be reduced towards zero as the size of the sub-group grows. In reality, however, not all ratings are reliably from $\mathcal{D}_p$, and the empirical average often gives a bad estimate.

In this work, we study robust estimation in such scenarios where we only have a partially-reliable dataset, which is also highly incomplete – each reviewer only provides information about a small subset of the papers. These scenarios also appear naturally in other contexts such as crowdsourcing (Vuurens et al., 2011) and peer-grading for online courses (Piech et al., 2013; Kulkarni et al., 2013), where small pieces of data are aggregated from a large population that cannot be completely trusted.

Recently, there has been a line of work on robustly estimating the mean and covariance of a distribution given a partially-reliable dataset (Diakonikolas and Kane, 2019; Diakonikolas et al., 2019a; Lai et al., 2016; Charikar et al., 2017; Diakonikolas et al., 2017a, 2018; Steinhardt et al., 2018; Cheng et al., 2019; Depersin and Lecué, 2019; Dong et al., 2019a; Lei et al., 2020; Cheng et al., 2020; Cherapanamjeri et al., 2020; Zhu et al., 2020a; Diakonikolas et al., 2020; Hopkins et al., 2020). Following the pioneering works by Tukey (1960) and Huber et al. (1964) in robust statistics, the dataset is modeled as initially i.i.d. examples $X^{(1)}, \ldots, X^{(N)} \in \mathbb{R}^d$ from the distribution, but εN of the examples are arbitrarily corrupted by an adversary. The presence of an ε-fraction of adversarial corruption not only makes the empirical average a bad estimator for the mean, but also makes it information-theoretically impossible to reduce the estimation error towards zero even with infinitely many examples. For estimating the mean of a spherical Gaussian, the Tukey median (Tukey, 1975) achieves minimax-optimal error guarantees (Chen et al., 2018; Zhu et al., 2020b), but it is hard to compute in general (Johnson and Preparata, 1978; Diakonikolas et al., 2017c); the linear-time-computable coordinate-wise median gives much worse L_2 errors that grow with the dimension. Recent breakthroughs by Diakonikolas et al. (2019a) and Lai et al. (2016) led to the best of both worlds: there are nearly-linear time algorithms achieving information-theoretically optimal error guarantees (Cheng et al., 2019; Depersin and Lecué, 2019; Dong et al., 2019a; Lei et al., 2020).

One limitation of these existing works is that they require observing all the d coordinates of each example, which is often impractical when the dimension is huge. In peer-grading, for example, a single student can only grade a very small fraction of the class. We thus ask:

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

can one estimate the mean of a d-dimensional distribution given an ε-corrupted dataset in which most of the coordinates of every example are missing?

If a particular coordinate is missing in every example, there is certainly no way to estimate that coordinate of the mean accurately. We thus assume that every coordinate appears in at least γN examples for $\gamma = \Omega(\varepsilon)$ so that the majority of the present coordinates are not corrupted. Even with this assumption, the incompleteness of the dataset brings additional challenges. The Tukey median, for example, is not well-defined in such an incomplete dataset. Natural approaches to eliminating the missing coordinates also fail as we show in Section 5. Nonetheless, we can still achieve optimal dimension-independent error guarantees in nearly-linear time as if there were no missing coordinates (see Section 1.2 for our results).

Following previous works (Diakonikolas et al., 2017a; Cheng et al., 2019; Dong et al., 2019a), we focus on two classes of distributions: 1) sub-Gaussian distributions with equal variance along all directions, and 2) distributions with bounded variances along all directions (see Section 2.1 for formal definitions). Assuming that the present non-corrupted coordinates are drawn from such a distribution, our algorithms estimate the mean of the distribution by starting with a coarse guess obtained by taking the coordinate-wise median, and refining the guess iteratively. In order to apply existing algorithms which only work for complete datasets, we first fill the missing coordinates with pre-determined values drawn i.i.d. from a mean-zero Gaussian distribution. In each iteration, we adjust the pre-determined values according to our current guess of the mean, and apply an existing algorithm to the resulting dataset to compute the next guess. We show that every new guess improves a potential function by a multiplicative factor, allowing us to eventually reduce the estimation error to the information-theoretical limit. The properties we need for the existing algorithm are described in Theorem 4, which we prove using the algorithm by Dong et al. (2019b) to obtain the best running time.

1.1 Corruption Model

We specify our assumptions on how the corrupted and incomplete dataset is generated. We assume that all the examples are drawn independently from "nice" distributions in a class $\mathcal{H}$ before some of them are corrupted. Specifically, we assume our dataset $X^{(1)}, \dots, X^{(N)} \in (\mathbb{R} \cup \{*\})^d$ is generated by an adversary according to the following procedure, where $*$ is the placeholder for the missing coordinates:

1. The adversary chooses a set of present (non-missing) coordinates $P^{(i)} \subseteq \{1, \dots, d\}$ for every $i = 1, \dots, N$.
2. The adversary chooses a d-dimensional distribution $\mathcal{D}^{(i)} \in \mathcal{H}$ for every $i = 1, \dots, N$, and draws $X^{(i)} \in \mathbb{R}^d$ independently from $\mathcal{D}^{(i)}$. Every $\mathcal{D}^{(i)}$ must have the same mean $\mu \in \mathbb{R}^d$, which is also chosen by the adversary.
3. The adversary modifies at most εN examples $X^{(i)}$ arbitrarily, potentially changing the corresponding $P^{(i)}$;
4. For every example $X^{(i)}$, its coordinates $X^{(i)}_j$ are concealed by $*$ whenever $j \notin P^{(i)}$.

In short, the adversary can choose for every example which coordinates are missing, and can corrupt an ε fraction of the examples after the examples are drawn. We call a dataset generated this way an *ε-corrupted dataset*. Figure 1 shows an example for $d = 4$ and $N = 7$.

$$
\begin{array}{l}
X^{(1)} \\ X^{(2)} \\ X^{(3)} \\ X^{(4)} \\ X^{(5)} \\ X^{(6)} \\ X^{(7)}
\end{array}
\begin{bmatrix}
1.2 & 1.8 & 2.9 & 4.0 \\
0.9 & 2.2 & 2.8 & 3.9 \\
0.8 & 1.9 & 3.1 & 4.1 \\
1.1 & 2.1 & 2.9 & 4.1 \\
1.0 & 2.0 & 3.0 & 3.8 \\
1.2 & 2.0 & 2.9 & 4.2 \\
1.2 & 2.1 & 3.2 & 3.9
\end{bmatrix}
$$

$$
\begin{array}{l}
X^{(1)} \\ X^{(2)} \\ X^{(3)} \\ X^{(4)} \\ X^{(5)} \\ X^{(6)} \\ X^{(7)}
\end{array}
\begin{bmatrix}
* & * & 2.9 & * \\
0.9 & * & 2.8 & 3.9 \\
* & * & * & 4.1 \\
\mathbf{0.5} & \mathbf{2.6} & * & \mathbf{5.0} \\
* & * & * & * \\
* & 2.0 & 2.9 & * \\
1.2 & 2.1 & * & *
\end{bmatrix}
$$

Figure 1: The orange entries are the present coordinates chosen in step 1. The top matrix is the dataset after step 2. The bottom matrix is what we observe after the adversary has corrupted $X^{(4)}$.

For technical convenience, we allow the adversary to be randomized. Thus, once the adversary is fixed (as a randomized algorithm), the randomness of the dataset comes from the random draws from $\mathcal{D}^{(i)}$ and the inherent randomness of the adversary. It is often more convenient to consider the present entries coordinate-wise rather than example-wise, so we define $\Gamma_j := \{i \in \{1, \dots, N\} : j \in P^{(i)}\}$ as the set of examples with coordinate j present. The adversary may sometimes choose Γ_j to be the empty set, in which case we have no hope to estimate μ_j. Therefore, we measure the performance of our algorithms only on γ-*complete datasets*, i.e., datasets with $|\Gamma_j| \geq \gamma N$ for all

coordinates j.

1.2 Our Results

Given an ε-corrupted dataset generated by the procedure in Section 1.1, we show nearly-linear time algorithms that compute good estimates for the true mean μ with high probability whenever the dataset is γ-complete. Specifically, we prove the following theorems for different classes of distributions defined more formally in Section 2.1:

Theorem 1. *Given an ε-corrupted dataset $X^{(1)}, \dots, X^{(N)} \in (\mathbb{R} \cup \{*\})^d$ generated with $\mathcal{H}$ being the class $\mathcal{H}_1(\eta)$ of 1-sub-Gaussian distributions with covariance $\eta^2 I$. Assume $C_0\varepsilon \le \gamma \le 1$, $\delta \in (0, 1/2)$, and $\gamma N \ge \Omega(\log(d/\delta))$, where $C_0 > 2$ is an absolute constant. There is an $\widetilde{O}(Nd\log(1/\delta))$-time algorithm with the following property: the event that the dataset is γ-complete but the algorithm does* not *output an estimate of the true mean μ up to L_2 error* err *happens with probability at* most *δ, where*

$$\mathsf{err} = O\Big(\sqrt{(d + \log(1/\delta))/\gamma N + (\varepsilon/\gamma)^2 \log(\gamma/\varepsilon)}\Big). \tag{1}$$

In particular, when

$$\gamma N \ge \Omega\left(\frac{\gamma^2(d + \log(1/\delta))}{\varepsilon^2 \log(\gamma/\varepsilon)}\right), \tag{2}$$

the error bound simplifies to

$$\mathsf{err} = O\Big((\varepsilon/\gamma)\sqrt{\log(\gamma/\varepsilon)}\Big). \tag{3}$$

Theorem 2. *Given an ε-corrupted dataset $X^{(1)}, \dots, X^{(N)} \in (\mathbb{R} \cup \{*\})^d$ generated with $\mathcal{H}$ being the class $\mathcal{H}_2$ of distributions with covariances $\Sigma \preceq I$. Assume $C_0\varepsilon \le \gamma \le 1$, $\delta \in (0, 1/2)$ and $\gamma N \ge \Omega(\log(d/\delta))$, where $C_0 > 2$ is an absolute constant. There is an $\widetilde{O}(Nd\log(1/\delta))$-time algorithm with the following property: the event that the dataset is γ-complete but the algorithm does* not *output an estimate of the true mean μ up to L_2 error* err *happens with probability at* most *δ, where*

$$\mathsf{err} = O\Big(\sqrt{d\log(d/\delta)/\gamma N + \varepsilon/\gamma}\Big). \tag{4}$$

In particular, when

$$\gamma N \ge \Omega\left(\frac{\gamma d\log(d/\delta)}{\varepsilon}\right), \tag{5}$$

the error bound simplifies to

$$\mathsf{err} = O\Big(\sqrt{\varepsilon/\gamma}\Big). \tag{6}$$

Tightness of bounds. Our error guarantees (3) and (6) are independent of the dimension d. If we focus on the "hardest" regime $\eta \ge \Omega(1)$ in Theorem 1, the error guarantees (3) and (6) are information-theoretically optimal up to constants even with infinitely many examples. This is most easily seen by a reduction from the complete-data setting. Given an ε'-corrupted dataset of N' examples without missing coordinates, we can simply add trivial examples with only missing coordinates to create an ε-corrupted γ-complete dataset of size N for $\varepsilon = \varepsilon'\gamma$ and $N = N'/\gamma$. The optimal error guarantees $O\big(\varepsilon'\sqrt{\log(1/\varepsilon')}\big)$ and $O\big(\sqrt{\varepsilon'}\big)$ for complete data[1] then translate exactly to (3) and (6), respectively. Similarly, in order to achieve the error bounds (3) and (6), our sample complexities (2) and (5) are optimal up to logarithmic factors. It is an interesting open question to achieve *sub-Gaussian rates* in Theorem 2, i.e., to replace $d\log(d/\delta)$ by $d+\log(1/\delta)$ (see Depersin and Lecué (2019); Lei et al. (2020); Diakonikolas et al. (2020) for some recent progress in the complete data setting).

Assumptions on distributions. By scaling the entire dataset, for any known σ^2, Theorem 1 generalizes to σ^2-sub-Gaussian distributions and Theorem 2 generalizes to distributions with covariances $\Sigma \preceq \sigma^2 I$. If one wants to apply Theorem 1 to distributions whose covariances are not $\eta^2 I$, but some other known diagonal matrix Σ, one only needs to scale each coordinate properly to make the covariance matrix equal to I. However, if Σ is not diagonal, i.e., the coordinates are correlated with each other, it is less clear how one could perform such a transformation. This is in contrast to the case without missing coordinates where any known covariance can be reduced to identity. We leave the interesting problem of handling correlated coordinates better in corrupted incomplete datasets to future work. Also, if the covariance is unknown to us, the adversary can easily hide the correlation between a pair of coordinates in our corruption model by never creating an example with both coordinates present, so robust covariance estimation on an incomplete dataset requires different assumptions, which would also be an interesting topic for future work.

How presence/missingness is determined. In our corruption model, the locations of the present co-

[1] Both error guarantees $O\big(\varepsilon'\sqrt{\log(1/\varepsilon')}\big)$ and $O\big(\sqrt{\varepsilon'}\big)$ are optimal even for one-dimensional data. The optimality of $O\big(\varepsilon'\sqrt{\log(1/\varepsilon')}\big)$ is shown by Diakonikolas et al. (2017b, Lemma 19). The optimality of $O\big(\sqrt{\varepsilon'}\big)$ follows easily from the indistinguishability of the following two distributions on ε'-corrupted datasets: the singleton distribution at 0, and the distribution over $\{0, 1/\sqrt{\varepsilon'}\}$ where the probability of getting $1/\sqrt{\varepsilon'}$ is $0.1\varepsilon'$.

ordinates are chosen by the adversary in step 1. In some applications, however, the algorithm itself can choose the locations: in the paper reviewing setting, the program committee can decide the assignment of papers to reviewers. In such cases, we can effectively eliminate the missing coordinates if we choose their locations appropriately (see Section 5.2 where we also show that choosing the locations uniformly at random does *not* give the "easiest" instances). Nevertheless, our results hold even when the locations are adversarially chosen, applying more broadly to cases where the locations of the present coordinates are restricted by, for example, conflict of interest and the lack of expertise of a reviewer in certain research areas. In the other direction, one may consider a stronger adversary where the first two steps are swapped: the adversary decides which coordinates are present *after* seeing the entire data matrix. In this case, the estimation error would grow with the dimension even without corruption: for example, assuming every entry in the matrix is i.i.d. from $\{0, a\}$ where the probability of getting 0 is slightly higher than γ, the adversary can choose to include i in Γ_j only when $X_j^{(i)} = 0$, making it impossible to accurately estimate the mean μ which clearly depends on a (see Liu et al. (2020b) for more results on datasets where missing entries are chosen by such stronger adversaries).

1.3 Other Related Works

Our work is close in spirit to Steinhardt et al. (2016), who also considered corrupted and incomplete datasets. They made generally weaker assumptions on the dataset: the majority of the examples may be corrupted, fewer coordinates are observed from each example, and different uncorrupted examples may come from different distributions as long as the discrepancies between pairs of coordinates are mostly preserved. Consequently, their goal is also weaker, which is to identify βd coordinates with approximately the highest average mean given the additional ability to obtain reliable values of a small number of coordinates.

Besides mean and covariance estimation, many other problems have been studied with corrupted data, such as principal component analysis (Candès et al., 2011; Chandrasekaran et al., 2011; Xu et al., 2010), learning graphical models Diakonikolas et al. (2016), linear regression (Bhatia et al., 2015, 2017; Diakonikolas et al., 2019b), sparse estimation (Balakrishnan et al., 2017; Liu et al., 2020a, 2019), and learning discrete distributions (Qiao and Valiant, 2018; Chen et al., 2020). More general models of data corruption are considered by Zhu et al. (2019), as well as in the robust hypothesis testing literature (Huber and Strassen, 1973; Verdu and Poor, 1984; Levy, 2008; Gül, 2017; Gül and Zoubir, 2017; Gao et al., 2018).

Early influential works on statistical inference using incomplete data include Rubin (1976), Dempster et al. (1977), and Rubin (1979). Data incompleteness is also a general theme of more recent research. For instance, matrix completion is the problem of recovering missing entries in a data matrix, usually under the assumption that the matrix is (approximately) low-rank (Candès and Recht, 2009; Candès and Tao, 2010; Candes and Plan, 2010). Trace reconstruction is the problem of reconstructing a sequence from its subsequences (traces) (Batu et al., 2004; Holenstein et al., 2008).

Even without data corruption or incompleteness, it is non-trivial to estimate the mean of a distribution with *optimal* error rates w.r.t. a small given failure probability bound δ, especially when the distribution is *heavy-tailed*. Lugosi et al. (2019) first showed that achieving *sub-Gaussian rates* is possible assuming only bounded second moment (see Lugosi and Mendelson (2019) for a more comprehensive survey).

2 PRELIMINARIES

2.1 Classes of Distributions

For $\sigma \geq 0$, we say a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ is σ^2-sub-Gaussian if for all vectors v, it holds that $\mathbb{E}_{X\sim\mathcal{D}}[\exp((X-\mu)^{\mathrm{T}}v)] \leq \exp(\sigma^2\|v\|_2^2/2)$. Given $\eta \geq 0$, class $\mathcal{H}_1(\eta)$ consists of 1-sub-Gaussian distributions with covariance $\eta^2 I$. Class $\mathcal{H}_2$ consists of distributions with covariances $\Sigma \preceq I$. It is easy to check that any σ^2-sub-Gaussian distribution has covariance $\Sigma \preceq \sigma^2 I$, so $\mathcal{H}_1(\eta)$ is non-empty only when $\eta \leq 1$. We thus assume $\eta \leq 1$ throughout the paper. The following simple property of $\mathcal{H}_1(\eta)$ and $\mathcal{H}_2$ turns out to be handy in the analysis of our algorithm – both $\mathcal{H}_1(\eta)$ and $\mathcal{H}_2$ are closed under coordinate-wise composition (see Appendix A for proof):

Claim 3. *Let $\mathcal{H}$ be $\mathcal{H}_1(\eta)$ or $\mathcal{H}_2$. Given m distributions $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(m)} \in \mathcal{H}$ all of which have the same mean μ, we define as follows their coordinate-wise composition $\mathcal{D}$ w.r.t. a partition $J^{(1)}, \ldots, J^{(m)}$ of the coordinates $\{1, \ldots, d\}$. We first draw $X^{(1)}, \ldots, X^{(m)}$ independently from $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(m)}$, respectively, and then construct a variable X so that it agrees with $X^{(i)}$ on the coordinates in $J^{(i)}$ for all i. We define the composition $\mathcal{D}$ to be the distribution of X. Then, $\mathcal{D}$ belongs to $\mathcal{H}$ and it has mean μ.*

2.2 Other Notations

For any $G \subseteq \{1, \ldots, N\}$, we define Δ_G as the set of weights $(w^{(1)}, \ldots, w^{(N)})$ satisfying $\forall i, w^{(i)} \geq 0$, $\sum_{i=1}^N w^{(i)} = 1$, and $\forall i \notin G, w^{(i)} = 0$. For any

$p \in (0, |G|]$, define $\Delta_{G,p}$ as the set of weights in Δ_G satisfying $\forall i, w^{(i)} \le \frac{1}{p}$. When p is an integer, $\Delta_{G,p}$ is the convex hull of uniform distributions over size-p subsets of G. We use $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{op}}$ to denote the vector L_2 norm and its induced matrix operator norm (a.k.a. the spectral norm), respectively.

3 ESTIMATION UNDER DETERMINISTIC CONDITIONS

Our goal is to extract useful information from a corrupted and highly incomplete data matrix: a fraction of the examples are controlled by an adversary, and most of the entries in our data matrix are missing. Somewhat surprisingly, given the terrible condition of the data matrix, some important structures are still preserved. In this section, we formally characterize the useful structures by a set of deterministic conditions, and show that we can perform efficient mean estimation as long as these deterministic conditions hold, regardless of how the data matrix was generated. In the next section (Section 4), we prove these deterministic conditions assuming the dataset is generated by the procedure in Section 1.1, and thus establish our main theorems.

3.1 A General Result Assuming No Missing Coordinates

When there are no missing coordinates, the deterministic conditions sufficient for performing robust mean estimation have been identified by many previous works (Diakonikolas et al., 2017a; Steinhardt et al., 2018; Cheng et al., 2019). We state such a set of conditions by defining the goodness property of a dataset (Definition 1), and show that this property is sufficient for robust mean estimation in Theorem 4, which summarizes the results by Dong et al. (2019b).

Definition 1 (summarized based on Cheng et al. (2019))**.** *We say a dataset $X^{(1)}, \dots, X^{(N)} \in \mathbb{R}^d$ is $(\varepsilon, \eta, \beta)$-good with respect to $\mu \in \mathbb{R}^d$ if there exists $G \subseteq \{1, \dots, N\}$ with the following properties:*

- $|G| \ge (1-\varepsilon)N$;
- *For all $w \in \Delta_{G,(1-3\varepsilon)N}$,*

$$\Big\| \sum_{i \in G} w^{(i)} (X^{(i)} - \mu) \Big\|_2 \le \beta\sqrt{\varepsilon}, \text{ and}$$
$$\Big\| \Big(\sum_{i \in G} w^{(i)} (X^{(i)} - \mu)(X^{(i)} - \mu)^{\mathrm{T}} \Big) - \eta^2 I \Big\|_{\mathrm{op}} \le \beta^2.$$

Theorem 4 (implicit in Dong et al. (2019b))**.** *There are absolute constants $c_1 \in (0, 1/3)$ and $C_2 > 0$ with the following property. Suppose that we have an $(\varepsilon, \eta, \beta)$-good dataset $X^{(1)}, \dots, X^{(N)} \in \mathbb{R}^d$ with respect to some unknown μ, and assume that $0 \le \varepsilon \le c_1, \beta^2 \ge \Omega(\eta^2 \varepsilon \log(1/\varepsilon)), \delta \in (0, 1/2)$. There is an $\widetilde{O}(Nd \log(1/\delta))$-time algorithm computing an estimate ν which, with probability at least $1-\delta$, satisfies that*

$$\|\nu - \mu\|_2^2 \le C_2 \beta^2 \varepsilon.$$

We defer the proof of Theorem 4 to Appendix B.

3.2 Improving the Guess When Coordinates are Missing

When our examples contain missing coordinates, the L_2 norm and the operator norm in Definition 1 are not well-defined. We address this issue by filling the missing coordinates according to a guess ν of the true mean μ, and prove Lemma 5 showing that the estimation error of ν controls the goodness of the resulting complete dataset, on which we can compute our next guess ν' using the algorithm from Theorem 4. In Theorem 6, we bound the estimation error of the new guess ν' directly by the estimation error of the old guess ν together with a goodness measure for the original *incomplete* dataset, which we define in Definition 2. This ensures that our guess becomes accurate enough for proving our main theorems after $O(\log N)$ iterations, as we will show in the next section (Section 4).

Given an incomplete dataset $X^{(1)}, \dots, X^{(N)}$, we first fill the missing coordinates in a simple pre-determined way, which we will specify in Section 4. For now, it suffices to assume that we have a *completed dataset* $X^{(1,0)}, \dots, X^{(N,0)} \in \mathbb{R}^d$ that agrees with the original dataset on the present entries. Given a guess ν, we define the *adjusted dataset* $X^{(1,\nu)}, \dots, X^{(N,\nu)}$ by

$$X_j^{(i,\nu)} = \begin{cases} X_j^{(i,0)} = X_j^{(i)}, & \text{if } X_j^{(i)} \ne *, \\ X_j^{(i,0)} + \nu_j, & \text{otherwise.} \end{cases} \tag{7}$$

We define the deterministic conditions on the completed dataset $X^{(1,0)}, \dots, X^{(N,0)}$:

Definition 2. *Given an incomplete dataset $X^{(1)}, \dots, X^{(N)} \in (\mathbb{R} \cup \{*\})^d$, recall that Γ_j is the set of examples with coordinate j present, i.e., $X_j^{(i)} \ne * \iff i \in \Gamma_j$. We say a corresponding completed dataset $X^{(1,0)}, \dots, X^{(N,0)} \in \mathbb{R}^d$ is $(\varepsilon, g_*, \eta, \beta)$-good with respect to μ if there exists $G \subseteq \{1, \dots, N\}$ with the following properties:*

- $|G| \ge (1-\varepsilon)N$;
- *$g_j := |G_j|/|G| \ge g_*$ for all $j = 1, \dots, d$, where $G_j := G \cap \Gamma_j$;*

- *For all* $w \in \Delta_{G,(1-3\varepsilon)N}$,

$$\Big\|\sum_{i\in G} w^{(i)}(X^{(i,\mu)}-\mu)\Big\|_2 \le \beta\sqrt{\varepsilon},$$
$$\Big\|\Big(\sum_{i\in G} w^{(i)}(X^{(i,\mu)}-\mu)(X^{(i,\mu)}-\mu)^{\mathrm{T}}\Big)-\eta^2 I\Big\|_{\mathrm{op}} \le \beta^2,$$
$$\Big\|\sum_{i\in G_j} w^{(i)}(X^{(i,\mu)}-\mu)\Big\|_2 \le \beta\sqrt{\varepsilon}, \quad \forall j=1,\dots,d.$$

We differentiate Definition 2 from Definition 1 by the number of parameters: Definition 2 has an additional parameter g_* suggesting the incomplete nature of the original dataset. It is clear that a completed dataset $X^{(1,0)},\dots,X^{(N,0)}$ being $(\varepsilon,g_*,\eta,\beta)$-good with respect to μ (Definition 2) implies that the adjusted dataset $X^{(1,\mu)},\dots,X^{(N,\mu)}$ being (ε,η,β)-good (Definition 1). However, this adjusted dataset depends on the unknown μ, so we cannot apply Theorem 4 directly to it. The following lemma (proved in Appendix C) shows that if we instead use a guess ν to adjust the dataset, the goodness of the resulting dataset is controlled by the estimation error of ν measured by

$$\|\nu-\mu\|_g := \Big(\sum_{j=1}^{d} g_j(\nu_j-\mu_j)^2\Big)^{1/2}.$$

Lemma 5. *There are absolute constants $C_3>1, C_4>1$ with the following property. Suppose we have an $(\varepsilon,g_*,\eta,\beta)$-good completed dataset $X^{(1,0)},\dots,X^{(N,0)}$ with respect to μ. Assume $5\varepsilon\le g_*\le 1$. Suppose $G,g_1,\dots,g_d$ satisfy the conditions in Definition 2. Let ν be a guess with $\|\nu-\mu\|_g\le\rho$. Then the adjusted dataset $X^{(1,\nu)},\dots,X^{(N,\nu)}$ is $(\varepsilon,\eta,\beta')$-good with respect to μ' where $\mu'_j := g_j\mu_j+(1-g_j)\nu_j$ and $(\beta')^2 := C_3\beta^2+C_4\rho^2$.*

Given the above goodness property of the adjusted dataset $X^{(1,\nu)},\dots,X^{(N,\nu)}$, we can apply the algorithm in Theorem 4 to it to compute an improved guess ν' (proved in Appendix D):

Theorem 6. *There are absolute constants $C_5\ge 5, C_6>0, c_7\in(0,1)$ with the following property. Suppose we have an $(\varepsilon,g_*,\eta,\beta)$-good completed dataset $X^{(1,0)},\dots,X^{(N,0)}\in\mathbb{R}^d$ with respect to μ, where $C_5\varepsilon\le g_*\le 1,\eta\ge 0,\beta^2\ge\Omega(\eta^2\varepsilon\log(1/\varepsilon))$. Suppose the $\mu,G,g_1,\dots,g_d$ satisfying the conditions in Definition 2 are all unknown. Given a current guess $\nu\in\mathbb{R}^d$, an upper bound ρ for $\|\nu-\mu\|_g$, and a desired failure probability bound $\delta\in(0,1/2)$, there is an $\widetilde{O}(Nd\log(1/\delta))$-time algorithm computing a new guess ν' which, with probability at least $1-\delta$, satisfies*

$$\|\nu'-\mu\|_g^2 \le C_6\beta^2\varepsilon+(1-c_7g_*)\rho^2.$$

4 FINITE SAMPLE CONCENTRATION

In this section, we prove Theorems 1 and 2 by showing that the deterministic conditions in Definition 2 are indeed satisfied by ε-corrupted γ-complete datasets generated according to Section 1.1 as long as we fill the missing entries properly to obtain the completed dataset: when $\mathcal{H}=\mathcal{H}_1(\eta)$, we fill the missing entries by independent $\mathcal{N}(0,\eta^2)$ variables, and when $\mathcal{H}=\mathcal{H}_2$, we fill the missing entries simply by zeros. In either case, we fill the missing entries using a mean-zero distribution from $\mathcal{H}$, so by Claim 3, when we adjust the completed dataset by the true mean μ, it would look like a dataset generated from $\mathcal{H}$ initially without missing entries. Specifically, we have the following two lemmas proved in Appendices E and F, respectively.

Lemma 7. *Assume an ε-corrupted dataset $X^{(1)},\dots,X^{(N)}\in\mathbb{R}^d$ is generated with $\mathcal{H}=\mathcal{H}_1(\eta)$. Assume $10\varepsilon\le\gamma\le 1,\delta\in(0,1/2)$. The event that the dataset is γ-complete but the corresponding completed dataset filled by independent $\mathcal{N}(0,\eta^2)$ variables is not $(\varepsilon',g_*,\eta,\beta)$-good happens with probability at most δ, where $\varepsilon'=\varepsilon, g_*=0.9\gamma$, and*

$$\beta=\Theta\Big(\sqrt{(d+\log(1/\delta))/N\varepsilon+\varepsilon\log(1/\varepsilon)}\Big).$$

Lemma 8. *Assume an ε-corrupted dataset $X^{(1)},\dots,X^{(N)}\in\mathbb{R}^d$ is generated with $\mathcal{H}=\mathcal{H}_2$. Assume $11\varepsilon\le\gamma\le 1,\delta\in(0,1/2)$, and $N\varepsilon\ge\Omega(\log(d/\delta))$. The event that the dataset is γ-complete but the corresponding completed dataset filled by zeros is not $(\varepsilon',g_*,\eta,\beta)$-good happens with probability at most δ, where $\varepsilon'=1.1\varepsilon, g_*=0.9\gamma,\eta=0$, and*

$$\beta=\Theta\Big(\sqrt{d\log(d/\delta)/N\varepsilon+1}\Big).$$

The goodness properties guaranteed by these lemmas allow us to apply Theorem 6 to improve our guess. We choose our initial guess as the coordinate-wise median with missing entries ignored, whose estimation error is bounded by the following claim (proved in Appendix G):

Claim 9. *Assume an ε-corrupted dataset $X^{(1)},\dots,X^{(N)}\in\mathbb{R}^d$ is generated with $\mathcal{H}$ being $\mathcal{H}_1(\eta)$ or $\mathcal{H}_2$. Assume $7\varepsilon\le\gamma\le 1,\delta\in(0,1/2),\gamma N\ge\Omega(\log(d/\delta))$. The event that the dataset is γ-complete but the coordinate-wise median ν does not satisfy $\|\nu-\mu\|_2\le O(\sqrt{d})$ happens with probability at most δ.*

We are now ready to summarize our algorithm and prove our main theorems:

Proof of Theorems 1 and 2. We will show a preprocessing procedure in the next section that allows

Algorithm 1: Robust mean estimation on incomplete data with arbitrary outliers

input : An ε-corrupted γ-complete dataset $X^{(1)}, \dots, X^{(N)}$ generated with $\mathcal{H}$ being $\mathcal{H}_1$ or $\mathcal{H}_2$.
output: An estimate ν of the mean μ.
Initialize ν as the coordinate-wise median;
Fill the missing entries as described to obtain a completed dataset $X^{(1,0)}, \dots, X^{(N,0)}$;
Define C_6, c_7 as in Theorem 6;
Define $\varepsilon', g_*, \eta, \beta$ as in Lemma 7 or Lemma 8 depending on whether $\mathcal{H} = \mathcal{H}_1(\eta)$ or $\mathcal{H} = \mathcal{H}_2$;
Initialize ρ as the upper bound for $\|\nu - \mu\|_2$ in Claim 9;
for $t = 1, \cdots, O(\log N)$ **do**
  Run the algorithm in Theorem 6 on the completed dataset with current guess ν, upper bound ρ, and parameters $\varepsilon', g_*, \eta, \beta$ to compute the next guess ν';
  $\nu \leftarrow \nu'$;
  $\rho \leftarrow (C_6\beta^2\varepsilon' + (1 - c_7 g_*)\rho^2)^{1/2}$;
end

us to assume w.l.o.g. that γ is lower-bounded by a small positive constant, i.e., $\gamma \geq \Omega(1)$.

Our assumption $\gamma N \geq \Omega(\log(d/\delta))$ implies that $\log(d/\delta)/\gamma N = O(1)$. We can then assume w.l.o.g. that $\varepsilon \geq \Omega(\log(d/\delta)/\gamma N)$ in Theorem 2 because for smaller ε, $d\log(d/\delta)/\gamma N$ dominates ε/γ in the error guarantee of Theorem 2. The assumption $N\varepsilon \geq \Omega(\log(d/\delta))$ in Lemma 8 is now satisfied.

We prove that Algorithm 1 satisfies the conditions of Theorems 1 and 2. The goodness property of the completed dataset $X^{(1,0)}, \dots, X^{(N,0)}$ is guaranteed by Lemma 7 and Lemma 8. By Theorem 6, in each iteration of Algorithm 1, when the current guess ν satisfies $\|\nu - \mu\|_g \leq \rho$, the next guess ν' satisfies $\|\nu' - \mu\|_g^2 \leq (\rho')^2 := C_6\beta^2\varepsilon' + (1 - c_7 g_*)\rho^2$, i.e.,

$$(\rho')^2 - \frac{C_6\beta^2\varepsilon'}{c_7 g_*} = (1 - c_7 g_*)\Big(\rho^2 - \frac{C_6\beta^2\varepsilon'}{c_7 g_*}\Big). \quad (8)$$

This shows a multiplicative decrease in a potential function. Since the initial ρ^2 is $O(d)$ given by Claim 9 and $g_* = 0.9\gamma \geq \Omega(1)$, in $O(\log(1 + d/\beta^2\varepsilon')) = O(\log N)$ iterations we obtain a guess ν^* with $\|\nu^* - \mu\|_2^2 \leq \|\nu^* - \mu\|_g^2/g_* \leq O(\beta^2\varepsilon')$, giving the desired error bounds for Theorems 1 and 2 after we plug in the values of ε', β from Lemma 7 and Lemma 8, respectively. The running time and the success probability of each iteration of Algorithm 1 are given by Theorem 6. □

5 SIMPLE BUT FAILED ATTEMPTS AND PRE-PROCESSING VIA HASHING

We discuss why two simpler methods (stacking and hashing) cannot directly reduce the robust mean estimation problem from the incomplete-data setting to the complete-data setting. We also show that we can use the hashing method as the pre-processing step in the proof of our main theorems in order to assume w.l.o.g. that γ is at least a constant.

5.1 Stacking

In a γ-complete dataset, each coordinate is present in at least γN examples. Thus, one may simply ignore the missing entries and stack the examples together to form a $\gamma N \times d$ matrix. For example, the dataset given in Figure 1 becomes

$$\begin{bmatrix} 0.9 & \mathbf{2.6} & 2.9 & 3.9 \\ \mathbf{0.5} & 2.0 & 2.8 & 4.1 \\ 1.2 & 2.1 & 2.9 & \mathbf{5.0} \end{bmatrix}.$$

Note that the corrupted entries (shown in bold) propagate to all the three resulting examples. In general, stacking breaks the fact that all the corruption happens in a small fraction of the examples, and this fact is essential for the L_2 error not to grow with the dimension.

5.2 Hashing

Stacking fails partly because different examples in the new dataset may depend on the same, possibly corrupted example in the original dataset. We thus hope that different new examples depend on *disjoint* subsets of the original examples. Imagine that step 1 of the corruption procedure in Section 1.1 is performed *not* by the adversary; instead, we have the freedom to choose which entries are present as long as no example has too many present coordinates, and the adversary is not allowed to modify $P^{(i)}$ in step 3. In this case, we can partition the N examples into $N' \approx \gamma N$ groups of similar sizes, and then for each group and each coordinate j, we choose an example in that group to have coordinate j present. In this way, the examples in each group can be combined into a new example *without missing coordinates* (see Figure 2). The j-th coordinate of the new example equals to $X_j^{(i)}$ where $X^{(i)}$ is an original example in the group with its j-th coordinate present. Hence, the new dataset contains no missing coordinates and has at most εN corrupted examples. Moreover, Claim 3 shows that the new dataset can be regarded as an $\frac{\varepsilon N}{N'} \approx \frac{\varepsilon}{\gamma}$-corrupted 1-complete dataset.

$$\begin{array}{c} \mu \\ X^{(1)} \\ X^{(2)} \\ \hline X^{(3)} \\ X^{(4)} \\ \hline X^{(5)} \\ X^{(6)} \\ X^{(7)} \end{array} \begin{array}{c} 1.0 \\ \left[\begin{array}{c} 1.2 \\ 0.9 \end{array}\right. \\ \hline \begin{array}{c} 0.8 \\ 1.1 \end{array} \\ \hline \begin{array}{c} 1.0 \\ 1.2 \\ 1.2 \end{array} \end{array} \begin{array}{c} 2.0 \\ \left[\begin{array}{c} 1.8 \\ 2.2 \end{array}\right. \\ \hline \begin{array}{c} 1.9 \\ 2.1 \end{array} \\ \hline \begin{array}{c} 2.0 \\ 2.0 \\ 2.1 \end{array} \end{array} \begin{array}{c} 3.0 \\ \left[\begin{array}{c} 2.9 \\ 2.8 \end{array}\right. \\ \hline \begin{array}{c} 3.1 \\ 2.9 \end{array} \\ \hline \begin{array}{c} 3.0 \\ 2.9 \\ 3.2 \end{array} \end{array} \begin{array}{c} 4.0 \\ \left[\begin{array}{c} 4.0 \\ 3.9 \end{array}\right. \\ \hline \begin{array}{c} 4.1 \\ 4.1 \end{array} \\ \hline \begin{array}{c} 3.8 \\ 4.2 \\ 3.9 \end{array} \end{array}$$

$$\begin{array}{c} \mu \\ X^{(1)} \\ X^{(2)} \\ \hline X^{(3)} \\ X^{(4)} \\ \hline X^{(5)} \\ X^{(6)} \\ X^{(7)} \end{array} \begin{array}{c} 1.0 \\ \left[\begin{array}{c} 1.2 \\ * \end{array}\right. \\ \hline \begin{array}{c} 0.8 \\ * \end{array} \\ \hline \begin{array}{c} 1.0 \\ * \\ * \end{array} \end{array} \begin{array}{c} 2.0 \\ \left[\begin{array}{c} 1.8 \\ * \end{array}\right. \\ \hline \begin{array}{c} * \\ \mathbf{2.6} \end{array} \\ \hline \begin{array}{c} 2.0 \\ * \\ * \end{array} \end{array} \begin{array}{c} 3.0 \\ \left[\begin{array}{c} * \\ 2.8 \end{array}\right. \\ \hline \begin{array}{c} 3.1 \\ * \end{array} \\ \hline \begin{array}{c} * \\ 2.9 \\ * \end{array} \end{array} \begin{array}{c} 4.0 \\ \left[\begin{array}{c} * \\ 3.9 \end{array}\right. \\ \hline \begin{array}{c} * \\ \mathbf{5.0} \end{array} \\ \hline \begin{array}{c} * \\ * \\ 3.9 \end{array} \end{array}$$

$$\begin{array}{c} \mu \\ \widetilde{X}^{(1)} \\ \widetilde{X}^{(2)} \\ \widetilde{X}^{(3)} \end{array} \begin{array}{c} 1.0 \\ \left[\begin{array}{c} 1.2 \\ 0.8 \\ 1.0 \end{array}\right. \end{array} \begin{array}{c} 2.0 \\ \left[\begin{array}{c} 1.8 \\ \mathbf{2.6} \\ 2.0 \end{array}\right. \end{array} \begin{array}{c} 3.0 \\ \left[\begin{array}{c} 2.8 \\ 3.1 \\ 2.9 \end{array}\right. \end{array} \begin{array}{c} 4.0 \\ \left[\begin{array}{c} 3.9 \\ \mathbf{5.0} \\ 3.9 \end{array}\right. \end{array}$$

Figure 2: If we choose the present entries as the orange ones in the left matrix, the three groups of the observed examples (the middle matrix) correspond to three new examples with no missing coordinates (the right matrix). Here, $X^{(4)}$ is corrupted, and so is $\widetilde{X}^{(2)}$.

This allows us to apply existing mean estimation algorithms to the new dataset.

To extend this reduction to cases where the present entries are chosen by the adversary, we need to figure out how to partition the original examples into groups. The N' groups of the original examples essentially correspond to a hash function $h : \{1, \cdots, N\} \to \{1, \cdots, N'\}$ mapping original examples to new examples. Each new example $\widetilde{X}^{(i')}$ is defined by $\widetilde{X}_j^{(i')} = X_j^{(i)}$ for some $i \in h^{-1}(i')$ satisfying $X_j^{(i)} \neq *$. For definiteness, let us always choose the smallest one if there are multiple such i, and define $\widetilde{X}_j^{(i')} = *$ if no such i exists. For example, if h maps $1, \cdots, 7$ into $2, 1, 3, 1, 1, 3, 2$, the dataset given in Figure 1 is transformed to

$$\begin{bmatrix} 0.9 & \mathbf{2.6} & 2.8 & 3.9 \\ 1.2 & 2.1 & 2.9 & * \\ * & 2.0 & 2.9 & 4.1 \end{bmatrix}.$$

Unfortunately, even if every coordinate j is present in γN examples chosen uniformly at random and independently for every j, one has to choose N' to be as small as $o(\gamma N)$ when $d, N \to +\infty$ in order to effectively eliminate the missing entries by hashing (see Claim 15 in Appendix H for details). In that case, the total number of present entries will reduce significantly in the new dataset. Moreover, if γ is only a constant factor larger than ε, the total number of new examples may become smaller than the number of corrupted original examples ($N' \leq o(\gamma N) < \varepsilon N$), and the majority of the new examples will likely have corrupted coordinates.

5.3 Pre-processing via Random Hashing

We show that, for a sufficiently small constant $c > 0$, we can assume without loss of generality that $\gamma \geq c$ in our main theorems (Theorems 1 and 2) by performing a pre-processing step. By "sufficiently small", we mean there is a large integer $B > 2$ such that $c \leq (B-2)/B^2$. If $\gamma < c$ instead, we hash the N examples into $N' := B\lceil \gamma N \rceil$ new examples as in Section 5.2. We use a random hash function h where $h(i)$ is independently uniformly chosen from $\{1, \cdots, N'\}$ for every $i = 1, \cdots, N$. There are at most $\varepsilon N \leq \frac{\varepsilon}{B\gamma} N'$ corrupted new examples, so Claim 3 guarantees that the new dataset is an ε'-corrupted dataset with the same μ for $\varepsilon' := \frac{\varepsilon}{B\gamma}$. To lower-bound the number of present coordinates in the new dataset, we note that $\widetilde{X}_j^{(i')}$ is present as long as any of the original examples $i \in \{1, \cdots, N\}$ with $X_j^{(i)} \neq *$ is hashed to i'. When the original dataset is γ-complete, the number of new examples with coordinate j present is at least the number of non-empty bins if we throw $\lceil \gamma N \rceil$ balls into $B\lceil \gamma N \rceil$ bins uniformly at random. If we throw the balls one by one, the conditional probability of each ball being thrown into an empty bin is at least $(B-1)/B$, so the number of non-empty bins is at least $((B-2)/B)\lceil \gamma N \rceil$ with probability $1 - \delta/2d$ by the Chernoff bound (recall our assumption $\gamma N \geq \Omega(\log(d/\delta))$). By a union bound over all the d coordinates, the new dataset is $\gamma' := ((B-2)/B)\lceil \gamma N \rceil / N' = (B-2)/B^2$-complete with probability at least $1 - \delta/2$ conditioned on the original dataset being γ-complete.

We can now run the algorithms in our main theorems on the newly created ε'-corrupted γ'-complete dataset of size N' where, indeed, $\gamma' = (B-2)/B^2 \geq c$. Note that the values γ/ε and γN barely change: $\gamma'/\varepsilon' = ((B-2)/B)\gamma/\varepsilon$ and $\gamma' N' \geq ((B-2)/B)\gamma N$. Therefore, all the assumptions of our main theorems still hold on the new dataset and our guarantees can translate back to the original dataset, up to tiny loss in the constants. The pre-processing step clearly runs in $O(Nd)$ time, and the size of the dataset does not increase.

6 CONCLUSIONS AND FUTURE RESEARCH

Motivated by applications such as crowdsourcing, we asked the natural theoretical question of robust mean estimation on incomplete data and solved it (nearly) optimally in terms of the error guarantee, the sample complexity, and the running time. As we discussed in Section 1.2, three natural questions remain unanswered and would be interesting topics for future research: 1) achieving sub-Gaussian rates in Theorem 2, 2) extending Theorem 1 to general known covariances, and 3) robustly estimating the covariance given incomplete data with adversarial corruption. Another interesting direction is to generalize our algorithms to settings where a majority of the present entries may be corrupted, i.e., $\gamma < 2\varepsilon$. This has been considered in the complete-data setting by Charikar et al. (2017) and Cherapanamjeri et al. (2020) using list-decoding, where the algorithm outputs a list of estimates with one of them likely being accurate. The list-decoding idea also applies when the underlying distribution is a mixture, where we want to estimate the mean of each component. Extending this list-decoding idea to incomplete data would lead to practical algorithms that can understand the views of diverse subgroups in the population from crowdsourced data. We hope that our iterative guessing-and-improving algorithm provides insights for answering these questions and for solving other problems in the incomplete-data setting.

Acknowledgements

We thank Michael P. Kim for valuable discussions at early stages of this work. We thank Jacob Steinhardt and Banghua Zhu for discussions on related research. We thank Hilal Asi, Mayee Chen, Dan Fu, Vatsal Sharan, and anonymous reviewers for helpful comments on earlier versions of this paper.

LH is supported by NSF Award IIS-1908774 and a VMware fellowship. OR is supported in part by NSF Award IIS-1908774 and by Simons Foundation investigators award 689988.

References

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212.

Batu, T., Kannan, S., Khanna, S., and McGregor, A. (2004). Reconstructing strings from random traces. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 910–918. ACM, New York.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2017). Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2110–2119.

Bhatia, K., Jain, P., and Kar, P. (2015). Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596.

Charikar, M., Steinhardt, J., and Valiant, G. (2017). Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60.

Chen, M., Gao, C., Ren, Z., et al. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960.

Chen, S., Li, J., and Moitra, A. (2020). Efficiently learning structured distributions from untrusted batches. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 960–973.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019). High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM.

Cheng, Y., Diakonikolas, I., Ge, R., and Soltanolkotabi, M. (2020). High-dimensional robust mean estimation via gradient descent. In *International Conference on Machine Learning*.

Cherapanamjeri, Y., Mohanty, S., and Yau, M. (2020). List decodable mean estimation in nearly linear time. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 141–148.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data

via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.

Depersin, J. and Lecué, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019a). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017a). Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017b). Robustly learning a gaussian: Getting optimal error, efficiently. *arXiv preprint arXiv:1704.03866*.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2018). Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Diakonikolas, I., Kane, D. M., and Pensia, A. (2020). Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2016). Robust learning of fixed-structure bayesian networks. *CoRR, abs/1606.07384*.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2017c). Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE.

Diakonikolas, I., Kong, W., and Stewart, A. (2019b). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM.

Dong, Y., Hopkins, S., and Li, J. (2019a). Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *Advances in Neural Information Processing Systems*, pages 6065–6075.

Dong, Y., Hopkins, S. B., and Li, J. (2019b). Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *arXiv preprint arXiv:1906.11366*.

Gao, R., Xie, L., Xie, Y., and Xu, H. (2018). Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912.

Gül, G. (2017). *Robust and distributed hypothesis testing*. Springer.

Gül, G. and Zoubir, A. M. (2017). Minimax robust hypothesis testing. *IEEE Transactions on Information Theory*, 63(9):5572–5587.

Holenstein, T., Mitzenmacher, M., Panigrahy, R., and Wieder, U. (2008). Trace reconstruction with constant deletion probability and related results. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 389–398. ACM, New York.

Hopkins, S. B., Li, J., and Zhang, F. (2020). Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33.

Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, pages 251–263.

Johnson, D. S. and Preparata, F. P. (1978). The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(6):1–31.

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

Lei, Z., Luh, K., Venkat, P., and Zhang, F. (2020). A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 2598–2612. PMLR.

Levy, B. C. (2008). Robust hypothesis testing with a relative entropy tolerance. *IEEE Transactions on Information Theory*, 55(1):413–421.

Li, J. Z. (2018). *Principled approaches to robust machine learning and beyond*. PhD thesis, Massachusetts Institute of Technology.

Liu, L., Li, T., and Caramanis, C. (2019). High dimensional robust m-estimation: Arbitrary corruption and heavy tails. *arXiv preprint arXiv:1901.08237*.

Liu, L., Shen, Y., Li, T., and Caramanis, C. (2020a). High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pages 411–421. PMLR.

Liu, Z., Park, J., Palumbo, N., Rekatsinas, T., and Tzamos, C. (2020b). Robust mean estimation under coordinate-level corruption with missing entries. *arXiv preprint arXiv:2002.04137*.

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

Lugosi, G., Mendelson, S., et al. (2019). Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794.

Maurer, A. et al. (2019). A Bernstein-type inequality for functions of bounded interaction. *Bernoulli*, 25(2):1451–1471.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*.

Qiao, M. and Valiant, G. (2018). Learning discrete distributions from untrusted batches. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Rubin, D. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. In *Proceedings of the 42nd session of the International Statistical Institute, Vol. 2 (Manila, 1979)*, volume 48, pages 517–532.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592. With comments by R. J. A. Little and a reply by the author.

Steinhardt, J., Charikar, M., and Valiant, G. (2018). Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Steinhardt, J., Valiant, G., and Charikar, M. (2016). Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. In *Advances in Neural Information Processing Systems*, pages 4439–4447.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485.

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

Verdu, S. and Poor, H. (1984). On minimax robustness: A general approach and applications. *IEEE Transactions on Information Theory*, 30(2):328–340.

Vuurens, J., de Vries, A. P., and Eickhoff, C. (2011). How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26.

Xu, H., Caramanis, C., and Mannor, S. (2010). Principal component analysis with contaminated data: The high dimensional case. *arXiv preprint arXiv:1002.4658*.

Zhu, B., Jiao, J., and Steinhardt, J. (2019). Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*.

Zhu, B., Jiao, J., and Steinhardt, J. (2020a). Robust estimation via generalized quasi-gradients. *arXiv preprint arXiv:2005.14073*.

Zhu, B., Jiao, J., and Steinhardt, J. (2020b). When does the Tukey median work? In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1201–1206. IEEE.