On Robust Mean Estimation under Coordinate-level Corruption

Zifan Liu [*1] **Jongho Park** [*1] **Theodoros Rekatsinas** [1] **Christos Tzamos** [1]

Abstract

We study the problem of robust mean estimation and introduce a novel Hamming distance-based measure of distribution shift for coordinate-level corruptions. We show that this measure yields adversary models that capture more realistic corruptions than those used in prior works, and present an information-theoretic analysis of robust mean estimation in these settings. We show that for structured distributions, methods that leverage the structure yield information theoretically more accurate mean estimation. We also focus on practical algorithms for robust mean estimation and study when data cleaning-inspired approaches that first fix corruptions in the input data and then perform robust mean estimation can match the information theoretic bounds of our analysis. We finally demonstrate experimentally that this two-step approach outperforms structure-agnostic robust estimation and provides accurate mean estimation even for high-magnitude corruption.

1. Introduction

Data corruption is an impediment to modern machine learning deployments. Corrupted data samples, i.e., data vectors with either noisy or missing values, can severely skew the statistical properties of a data set, and hence, lead to invalid inferences. Robust statistics seek to provide methods for problems such as estimating the mean and covariance of a data set that are resistant to data corruptions.

Much of the existing robust estimation methods assume that a data sample is either completely clean or completely corrupted; *Huber contamination model* (Huber, 1992) and the *strong contamination model* considered by Diakonikolas & Kane (2019) are typical examples of such corruption models. Under this kind of model, robust estimators rely either on filtering or down-weighting corrupted data vectors to reduce their influence (Diakonikolas et al., 2019a; 2017). In many applications, however, we can have partially corrupted data samples and even all data vectors can be partially corrupted. For example, in DNA microarrays, measurement errors or dropouts can occur for batches of genes (Troyanskaya et al., 2001b). Filtering or down-weighting an entire data vector can waste the information contained in the clean coordinates of the vector.

Moreover, recent works (Rekatsinas et al., 2017; Wu et al., 2020; Khosravi et al., 2019) show that to obtain state-of-the-art empirical results for predictive tasks over noisy data, one needs to leverage the redundancy in data samples introduced by statistical dependencies among the coordinates of the data (referred to as *structure* hereafter) to learn an accurate distribution of the clean data and use that to repair corruptions. This work aims to promote theoretical understanding as to why leveraging statistical dependencies is key in dealing with data corruption. To this end, we study the connections between robust statistical estimation under worst-case (e.g., adversarial) corruptions and *structure-aware* recovery of corrupted data.

Problem Summary We consider *robust mean estimation* under *coordinate-level* corruptions (either missing entries or value replacements). We study worst-case, adversarial corruption, i.e., we assume that corruption is systematic and cannot be modeled as random noise. We consider the *adversarial model* for which a given data set generated from an unknown distribution can have up to α-fraction of its coordinates corrupted adversarially, i.e., the adversary can strategically hide or modify individual coordinates of samples. The goal is to find an estimate $\hat{\mu}$ of the true mean μ of the data set that is accurate even in the worst case.

Main Contributions First, we present a new information theoretic analysis of robust mean estimation for coordinate-level corruptions, i.e., both replacement-based corruptions and missing-data corruptions. Our analysis introduces a model for coordinate-level corruptions, and d_{ENTRY}, a new measure of distribution shift for coordinate-level corruption. We base d_{ENTRY} on the Hamming distance between samples from the original and the corrupted distribution. The reason is that the Hamming distance between the original and the corrupted samples is at most the amount of corruption in

[*]Equal contribution [1]Department of Computer Science, University of Wisconsin-Madison, Madison, USA. Correspondence to: Zifan Liu <zliu676@wisc.edu>, Jongho Park <jongho.park@wisc.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

this sample, neatly reflecting the noise level.

We present an information-theoretic analysis of corruption under coordinate-level adversaries and formally validate the empirical observations in the data cleaning literature: one must exploit the structure of the data to achieve information-theoretically optimal error for mean estimation. To show that structure is key, we focus on the case where the data lies on a lower-dimensional subspace, i.e., the observed sample before corruption is $x = Az$, where $A \in \mathbb{R}^{n\times r}$ and $z \in \mathbb{R}^r$ is a lower-dimensional vector drawn from an unknown distribution D_z. Also, n is the total number of coordinates in x. Such low-rank subspace-structure is common in real-world data and the linear assumption is standard in theoretical exploration. We identify a key quantity m_A, the minimum number of rows that one needs to remove from A to reduce its row space by one, which captures the effect of structure on mean-estimation error $||\mu - \hat{\mu}||$. For Gaussian distributions, the de facto distribution considered in the robust statistics literature, we prove that no algorithm can achieve error better than $\Omega(\alpha\frac{n}{m_A})$ when α-fraction of coordinates per sample on average is adversarially corrupted. Our analysis highlights that, for coordinate-level corruption, it is necessary to use the structure in the data to perform recovery before statistical estimation. Specifically, when α-fraction of the values are corrupted, recently-introduced estimators (Diakonikolas et al., 2019a) yield an estimation error of $\Omega(\alpha n)$ which is not the information theoretic optimal in the presence of structure. We show that to achieve the information theoretic optimal error of $\Theta(\alpha\frac{n}{m_A})$, one needs to consider the dependencies amongst coordinates.

Second, we study the existence of practical algorithms with polynomial complexity which yield results that match the error bounds of our information-theoretic analysis. We first show that, when corruptions correspond to missing data, a data-cleaning-inspired two-step approach achieves the information-theoretic optimal error. Specifically, to achieve the optimal error, one must first use imputation strategies that leverage the dependencies between data coordinates to recover the missing entries and then proceed with robust mean estimation over the imputed data. We show that in the case of linear structure, if the dependencies across coordinates are modeled via a known matrix A, we can recover missing entries by solving a linear system; when A is unknown, we show that under bounded amount of corruption, we can leverage matrix completion methods (Pimentel-Alarcón et al., 2016) to recover the missing entries and obtain the same performance as in the case with known structure. We also explore replacement-type corruptions. By drawing connections to sparse recovery, we show that recovery for replacements is computationally intractable in general. As a preliminary result, we propose a randomized recovery algorithm for replacements that achieves probabilistic guarantees when the structure A is known.

Finally, we present an experimental evaluation of the aforementioned two-step approach for missing-data corruptions on real-world data and show it leads to significant accuracy improvements over prior robust estimators even for samples that do not follow a Gaussian distribution or whose structure does not conform to a linear model.

2. Background and Motivation

We review the problem of robust mean estimation and discuss models and measures related to our study.

Robust Mean Estimation Robust mean estimation seeks to recover the mean $\mu \in \mathbb{R}^n$ of a n-dimensional distribution D from a list of i.i.d. samples where an unknown number of arbitrary corruptions has been introduced in the samples. Given access to a collection of N samples $x_1, x_2, \ldots, x_N$ from D on $\mathbb{R}^n$ when a fraction of them have been fully or partially corrupted, robust mean estimation seeks to find a vector $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|$ is as small as possible. We consider two norms to measure the mean estimation error. The first norm is the Euclidean (ℓ_2) distance and the second is the scale-invariant *Mahalanobis distance* defined as $\|\mu - \hat{\mu}\|_\Sigma = |(\mu-\hat{\mu})^T\Sigma^{-1}(\mu-\hat{\mu})|^{1/2}$, where Σ is the covariance matrix. When the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance.

Sample-level Corruption A typical model to describe worst-case corruptions is that of a *sample-level adversary*, hereafter denoted $\mathcal{A}_1^\epsilon$. In this paper, we assume that this adversary is allowed to inspect the samples and corrupt an ϵ-fraction of them in an arbitrary manner. All coordinates of those samples are considered corrupted. Corruptions introduce a shift of the distribution D, which we can measure using the *total variation distance* (d_{TV}). Total variation distance between two distributions P and Q on $\mathbb{R}^n$ is defined as $d_{\text{TV}}(P,Q) = \sup_{E\subseteq\mathbb{R}^n} |P(E) - Q(E)|$ or equivalently $\frac{1}{2}\|P - Q\|_1$. For two Gaussians $D_1 = \mathcal{N}(\mu_1, \Sigma)$ and $D_2 = \mathcal{N}(\mu_2, \Sigma)$ with $d_{\text{TV}}(D_1, D_2) = \epsilon < 1/2$ it is that $\|\mu_1 - \mu_2\|_\Sigma = \Theta(\epsilon)$, i.e., their total variation distance and the Mahalanobis distance of their means are equivalent up to constants. This result allows tight analyses of Gaussian mean estimation for a bounded fraction of corruptions.

Motivation Total variation only provides a coarse measure of distribution shift, and hence, leads to a coarse characterization of the mean estimation error. For example, corruption of one coordinate per sample versus corruption of all coordinates results in the same distribution shift under total variation. However, the optimal mean estimation error can be different for these two cases. For example, if a corrupted coordinate has identical copies in other uncorrupted coordinates, the effect to mean estimation should be zero as we can repair the corrupted coordinate. This motivates our study.

3. Information Theoretic Analysis

We study robust mean estimation under fine-grained corruption schemes. First, we introduce two new coordinate-level corruption adversaries (models) and a new measure of distribution shift (d_{ENTRY}) that characterizes the effect of those adversaries on the observed distribution. Second, we present an information theoretic analysis of robust mean estimation and prove information-theoretically optimal bounds for mean estimation over Gaussians $\mathcal{N}(\mu, \Sigma)$ under coordinate-level corruption with respect to Mahalanobis distance. The results in this section hold for replacement-based corruptions as well as missing values. All proofs are deferred to the supplementary material of our paper.

3.1. Coordinate-level Corruption Adversaries

We introduce two new adversaries and compare them to the sample-level adversary $\mathcal{A}_1^\epsilon$ from Section 2:

First, we consider an extension of $\mathcal{A}_1^\epsilon$ to coordinates, and define a *value-fraction adversary*, denoted by $\mathcal{A}_2^\rho$. Given N samples from distribution D on $\mathbb{R}^n$, adversary $\mathcal{A}_2^\rho$ is allowed to corrupt up to a ρ-fraction of values in each coordinate of the N samples. This adversary can corrupt a total of $\rho \cdot N \cdot n$ values in the N samples; these values can be distributed strategically across samples leading to cases where *most* of the samples are corrupted but still the corruption per coordinate is bounded by ρN.

Second, we define the more powerful *coordinate-fraction adversary* $\mathcal{A}_3^\alpha$ that can corrupt *all samples* in the worst case. $\mathcal{A}_3^\alpha$ is allowed to corrupt up to α-fraction of all values in the N samples, i.e., up to a total of $\alpha \cdot N \cdot n$ values. When $\alpha \geq \frac{1}{n}$, adversary $\mathcal{A}_3^\alpha$ can corrupt *all* N samples.

Note that similar to $\mathcal{A}_1^\epsilon$, the coordinate-level adversaries we consider ($\mathcal{A}_2^\rho$, and $\mathcal{A}_3^\alpha$) are adaptive, i.e., they are allowed to inspect the samples before choosing a fraction of the coordinates to corrupt.

Adversary Comparison $\mathcal{A}_1^\epsilon$ corresponds to the standard adversary associated with the strong contamination model considered by Diakonikolas & Kane (2019), which either corrupts a sample completely or leaves it intact. Adversaries $\mathcal{A}_2^\rho$ and $\mathcal{A}_3^\alpha$ are more fine-grained since they can corrupt only part of the entries of a sample. As a result, $\mathcal{A}_2^\rho$ and $\mathcal{A}_3^\alpha$ can corrupt more samples than $\mathcal{A}_1^\epsilon$ for similar budget-fractions ϵ, ρ, and α.

We formalize the comparison among $\mathcal{A}_1^\epsilon$, $\mathcal{A}_2^\rho$, and $\mathcal{A}_3^\alpha$ in the next propositions. We seek to understand when an adversary $\mathcal{A}$ can *simulate* another $\mathcal{A}'$, i.e., $\mathcal{A}$ can perform any corruption performed by $\mathcal{A}'$.

Proposition 1. *If $\alpha,\ \rho \leq \epsilon/n$, then $\mathcal{A}_1^\epsilon$ can simulate $\mathcal{A}_2^\rho$ and $\mathcal{A}_3^\alpha$. If $\alpha \leq \rho/n$, $\mathcal{A}_2^\rho$ can simulate $\mathcal{A}_3^\alpha$.*

Proposition 2. *If $\alpha,\ \rho \geq \epsilon$, then $\mathcal{A}_2^\rho$ and $\mathcal{A}_3^\alpha$ can simulate $\mathcal{A}_1^\epsilon$. If $\alpha \geq \rho$, $\mathcal{A}_3^\alpha$ can simulate $\mathcal{A}_2^\rho$.*

These propositions show that the two adversary types (sample- and coordinate-level) can simulate each other under different budget conditions, thus, enabling reductions between the two types.

Proposition 1 implies that we can reduce coordinate-level corruption to sample-level corruption by considering $\mathcal{A}_3^\alpha$ as $\mathcal{A}_1^\epsilon$ with $\epsilon = \alpha n$. This reduction guarantees that *any algorithm for mean estimation with guarantees for $\mathcal{A}_1^\epsilon$ enjoys the same guarantees for coordinate-level corruption when $\epsilon \geq \alpha \cdot n$.* Similarly, Proposition 2 means that *any lower-bound guarantee on mean estimation for $\mathcal{A}_1^\epsilon$ also holds for $\mathcal{A}_3^\alpha$ when $\epsilon = \alpha$.* However, this characterization is loose as the gap between α and αn is large, raising the question: Are there distributions for which this gap is more tight and are there data properties we can exploit to reduce the dimensional factor of n? Next, we show that structure in data affects the power of coordinate-level corruption and introduces information-theoretically tight bounds for mean estimation under coordinate-level corruption.

3.2. Distribution Shift in Coordinate-level Corruption

We propose a new type of distribution shift metric, referred to as d_{ENTRY}, which can capture fine-grained coordinate-level corruption. We have the next definition:

Definition 1 (d_{ENTRY})**.** *Consider the coupling γ of two distributions P, Q, i.e., a joint distribution of P and Q such that the marginal distributions are P, Q. Let the set of all couplings of P, Q be $\Gamma(P, Q)$, and define for $x, y \in \mathbb{R}^n$, $I(x, y) = [\mathbb{1}_{x_1 \neq y_1}, \ldots, \mathbb{1}_{x_n \neq y_n}]^\top$. For D_1, D_2 on $\mathbb{R}^n$,*

$$d^1_{ENTRY}(D_1, D_2) = \inf_{\gamma \in \Gamma(D_1, D_2)} \frac{1}{n} \| \mathop{\mathbb{E}}_{(x,y)\sim\gamma} [I(x, y)] \|_1$$

$$d^\infty_{ENTRY}(D_1, D_2) = \inf_{\gamma \in \Gamma(D_1, D_2)} \| \mathop{\mathbb{E}}_{(x,y)\sim\gamma} [I(x, y)] \|_\infty$$

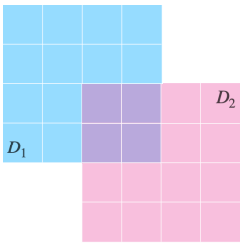
The following theorem shows the relation between d^1_{ENTRY} and $\mathcal{A}_3^\alpha$.

Theorem 1. *Let D_1, D_2 be two distributions such that $d^1_{ENTRY}(D_1, D_2) = \alpha'$. $\mathcal{A}_3^\alpha$ corrupts α fraction of N samples from D_1. If $\alpha > 2\alpha'$, $\mathcal{A}_3^\alpha$ has a way to make corruptions so that with probability at least $1 - e^{-\Omega(\alpha^2 N)}$ it is indistinguishable whether the N samples come from D_1 or D_2. If $\alpha < \alpha'/4$, no matter how $\mathcal{A}_3^\alpha$ makes corruptions, with probability at least $1 - e^{-\Omega(\alpha^2 N)}$, we can tell that the N samples come from D_1.*

The relation above also holds for d^∞_{ENTRY} and $\mathcal{A}_2^\rho$. The theorem shows that d_{ENTRY} gives a tight asymptotic bound on the power of coordinate-level adversaries.

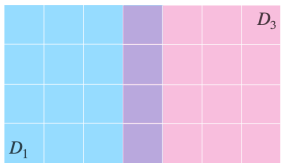
Intuitively $d^1_{\text{ENTRY}}(D_1, D_2)$ represents how many coordinates need to be corrupted (out of n on average) for D_1

$$d_{TV}(D_1, D_2) = \frac{3}{4}$$
$$d^1_{ENTRY}(D_1, D_2) = \frac{1}{2}$$



(a) Corruption of both coordinates

$$d_{TV}(D_1, D_3) = \frac{3}{4}$$
$$d^1_{ENTRY}(D_1, D_3) = \frac{3}{8}$$



(b) Corruption of one coordinate

Figure 1. Comparing d_{TV} and d_{ENTRY}: D_1 is the original 2D uniform distribution. D_2 is obtained after corruptions from $\mathcal{A}_3^\alpha$ with $\alpha = 1/2$, and D_3 after corruption from $\mathcal{A}_3^\alpha$ with $\alpha = 3/8$.

and D_2 to be indistinguishable. Then, given the original distribution D and sufficiently large sample size, $\{D' : d^1_{\text{ENTRY}}(D, D') \leq \alpha\}$ represents the set of distributions that $\mathcal{A}_3^\alpha$ can show us after corruption, and thus d^1_{ENTRY} allows us to capture all possible actions of this adversary. Similarly, d^∞_{ENTRY} captures all possible actions of $\mathcal{A}_2^\rho$. We use d_{ENTRY} when both d^1_{ENTRY} and d^∞_{ENTRY} apply.

We compare d^1_{ENTRY} and d_{TV} in Figure 1. We consider a 2D uniform distribution D_1 and two corrupted versions D_2 and D_3. D_2 is obtained after an adversary corrupts both coordinates for samples from the upper-left quadrant of D_1 and one of the coordinates for samples in the lower-left and upper-right quadrant of D_1. D_3 is obtained after an adversary corrupts the horizontal coordinate for samples obtained from the left-most $3/4$ of D_1. In both cases, $d_{\text{TV}}(D_1, D_2) = d_{\text{TV}}(D_1, D_3) = 3/4$ since $3/4$ of the samples from D_1 are corrupted. d_{ENTRY} is different: From Definition 1, we have $d^1_{\text{ENTRY}}(D_1, D_2) = 1/2$ and $d^1_{\text{ENTRY}}(D_1, D_3) = 3/8$, thus, we can distinguish the two.

3.3. Information-theoretic Bounds for Gaussians

We analyze robust mean estimation under coordinate-level corruptions for Gaussian distributions, the de facto choice in the robust estimation literature. This choice enables us to draw comparisons to prior mean estimation approaches. Our results are summarized in Table 1. We first show an impossibility result for arbitrary Gaussian distributions: in the general case, the information-theoretic analysis based on d_{TV} and sample-level adversaries (Tukey, 1975; Diakonikolas et al., 2019a) is tight even for coordinate-level corruption

Table 1. Our results for robust mean estimation ($||\hat{\mu} - \mu||_\Sigma$) under $\mathcal{A}_1^\epsilon$ (sample-level adversary), $\mathcal{A}_2^\rho$, and $\mathcal{A}_3^\alpha$ (coordinate-level adversaries). The results are for Gaussian distributions.

Structure	$\mathcal{A}_1^\epsilon$	$\mathcal{A}_2^\rho$	$\mathcal{A}_3^\alpha$
No Structure	$\Theta(\epsilon)$	$\Omega(\rho\sqrt{n}), O(\rho n)$	$\Theta(\alpha n)$
Linear structure A	$\Theta(\epsilon)$	$O(\rho\frac{n}{m_A})$	$\Theta(\alpha\frac{n}{m_A})$

adversaries. However, we show that this result *does not hold* for distributions that exhibit structure, i.e., redundancy across coordinates. We show that, for structured Gaussian distributions and corruptions that lead to a d_{ENTRY}-bounded distribution shift, one must exploit the structure to achieve information-theoretically optimal error for mean estimation.

Mean Estimation of Arbitrary Gaussians We consider a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with full rank covariance matrix Σ. We assume that observed samples are corrupted by a coordinate-level adversary. We first present a common upper-bound on the mean estimation error for both $\mathcal{A}_2^\rho$ and $\mathcal{A}_3^\alpha$, and then introduce the corresponding lower-bounds.

We obtain an upper-bound on $\|\hat{\mu} - \mu\|_\Sigma$ by using Proposition 1: A sample-level adversary can simulate a coordinate-level adversary when $\epsilon = \alpha \cdot n$. But, for Adversary $\mathcal{A}_1^\epsilon$ the Tukey median achieves optimal error $\|\hat{\mu} - \mu\|_\Sigma = \Theta(\epsilon)$ when $\epsilon < 1/2$. Thus, the Tukey median yields error $O(\alpha n)$ for the coordinate-level adversaries $\mathcal{A}_2^\rho$ (when $\rho = \alpha$) and $\mathcal{A}_3^\alpha$, when $\alpha \cdot n < 1/2$. Note that the condition $\alpha \cdot n < 1/2$ is necessary for achieving such an upper bound. When $\alpha \cdot n \geq 1/2$, the coordinate-level adversary is able to corrupt more than half of the samples in the worst case, leading to unbounded error, which shows exactly the power of the coordinate-level adversary.

We now focus on lower-bounds for the mean estimation error. We first consider adversary $\mathcal{A}_2^\rho$ who can corrupt at most ρ-fraction of each coordinate in the samples. For this setting, the optimal estimation error depends on the *disc* of the covariance matrix Σ, where disc is defined as:

Definition 2. *(disc) For a positive semi-definite matrix M, define $s(M)_{ij} = M_{ij}/\sqrt{M_{ii}M_{jj}}$ and disc$(M) = \max_{x\in[-1,1]} \sqrt{x^T s(M) x}$.*

Theorem 2. *Let $\Sigma \in \mathbb{R}^{n\times n}$ be full rank. Given a set of i.i.d. samples from $\mathcal{N}(\mu, \Sigma)$ where the set is corrupted by $\mathcal{A}_2^\rho$, any algorithm for estimating $\hat{\mu}$ must satisfy $\|\mu - \hat{\mu}\|_\Sigma = \Omega(\rho \cdot disc(\Sigma^{-1}))$.*

From this theorem, we obtain the next corollary for the mean estimation error for $\mathcal{A}_2^\rho$:

Corollary 1. *Given a set of i.i.d. samples from $\mathcal{N}(\mu, \Sigma)$ where the set is corrupted by $\mathcal{A}_2^\rho$, any algorithm that outputs a mean estimator $\hat{\mu}$ must satisfy $\|\mu - \hat{\mu}\|_\Sigma = \Omega(\rho\sqrt{n})$.*

We see that there is a gap between the lower and upper bound on the mean estimation error for $\mathcal{A}_2^\rho$. However, we show that such a gap does not hold for $\mathcal{A}_3^\alpha$. For $\mathcal{A}_3^\alpha$, the lower bound is the same as the upper bound presented above. Specifically, for the coordinate-fraction adversary $\mathcal{A}_3^\alpha$, it is impossible to achieve a mean estimation error better than $O(\alpha n)$ in the case of arbitrary Gaussian distributions:

Theorem 3. *Let $\Sigma \in \mathbb{R}^{n\times n}$ be full rank. Given a set of i.i.d. samples from $\mathcal{N}(\mu, \Sigma)$ where the set is corrupted by $\mathcal{A}_3^\alpha$, any algorithm that outputs a mean estimator $\hat{\mu}$ must satisfy $\|\hat{\mu} - \mu\|_\Sigma = \Omega(\alpha n)$.*

To gain some intuition, consider $\mathcal{A}_3^\alpha$ with $\alpha \geq \frac{1}{n}$. In this case, $\mathcal{A}_3^\alpha$ can concentrate all corruption in the first coordinate of all samples, and hence, we cannot estimate the mean for that coordinate. An immediate result is that for worst-case coordinate-corruptions, i.e., corruptions introduced by $\mathcal{A}_3^\alpha$, over arbitrary Gaussian distributions the mean estimation error is precisely $\|\mu - \hat{\mu}\|_\Sigma = \Theta(\alpha n)$.

Mean Estimation of Structured Gaussians The previous analysis for $\mathcal{A}_3^\alpha$ shows that we cannot improve upon existing algorithms. However, real-world data often exhibit structural relationships between features such that one may be able to infer corrupted values via other visible values (Wu et al., 2020). We show that in the presence of *structure* due to dependencies, one must exploit the structure of the data to achieve information-theoretically optimal error for mean estimation. To show that structure is key, we focus on samples $x_i \in \mathbb{R}^n$ that lie in a low-dimensional subspace such that $x_i = Az_i$, where $A \in \mathbb{R}^{n\times r}$ represents the structure. Such low-rank subspace-structure is natural in many real-world scenarios and we assume linearity for the convenience of analysis. In fact, linear structure can also encode more complex structures (e.g., polynomials) if one considers an augmented set of features. We assume z_i comes from a non-degenerate Gaussian in $\mathbb{R}^r$. We consider a data sample $x = Az$ before corruption and assume that corruption is introduced in x.

In this setting, the coordinate-level adversary has limited effect in mean estimation due to the redundancy that A introduces. We can measure the strength of this redundancy with respect to coordinate-level corruption by considering its row space. The coordinates of $x = Az$, and hence, the corrupted data, exhibit high redundancy when many rows of A span a small subspace. We define a quantity m_A to derive information-theoretic bounds for structured Gaussians.

Definition 3 (m_A)**.** *Given matrix $A \in \mathbb{R}^{n\times r}$, m_A is the minimum number of rows one needs to remove from A to reduce the dimension of its row space by one.*

When $A = I$ is the identity matrix, it is $m_I = 1$ and we can remove any row to reduce its row space; we have low redundancy. But, for $A = [\mathbf{e}_1, \ldots, \mathbf{e}_1]^\top$ where $\mathbf{e}_1$ has 1 in its first coordinate and 0 in the others, $m_A = n$ since we need to remove all $\mathbf{e}_1$'s to reduce A's row space. It holds that $1 \leq m_A \leq n$.

We next show that the higher the value that m_A takes, the weaker a coordinate-level adversary becomes due to the increased redundancy. Intuitively, the coordinate-level adversary has to spend more budget per sample to introduce corruptions that will counteract the redundancy introduced by matrix A. Theorem 4 shows that $\mathcal{A}_2^\rho, \mathcal{A}_3^\alpha$ cannot alter the original distribution too far in d_{TV}, leading to information-theoretically tight bounds for mean estimation.

Theorem 4. *Given two probability distributions D_1, D_2 on $\mathbb{R}^n$ with support in the range of linear transformation A,*

$$(m_A/n) \cdot d_{TV}(D_1, D_2) \leq d_{ENTRY}(D_1, D_2) \leq d_{TV}(D_1, D_2)$$

Here, D with support in the range of A means a distribution D that is generated such that it lies on the subspace generated by A, i.e., there is zero measure outside of this subspace. Since d_{TV} between two Gaussians is asymptotically equivalent to the Mahalanobis distance between them, we get the following corollary using d_{ENTRY}.

Corollary 2. *Let $\mathcal{N}(\mu, \Sigma)$ be a Gaussian with support in the range of linear transformation A. For $\hat{\mu}$ such that $d_{ENTRY}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \Sigma)) \leq \alpha$, $\|\mu - \hat{\mu}\|_\Sigma = O(\alpha \frac{n}{m_A})$.*

Note that the upper bound above is under the condition that $\alpha < m_A/(2n)$ when the corruption is limited to missing entries, and $\alpha < m_A/(4n)$ when replacement is allowed. Otherwise, more than half of the samples can be corrupted and unrecoverable (the proof of Theorem 5 provides the conditions for recovery, and the break points for mean estimation follow). Corollary 2 shows that $\mathcal{A}_2^\rho$ (when $\rho = \alpha$) and $\mathcal{A}_3^\alpha$ can only shift structured distributions by $O(\alpha \frac{n}{m_A})$. This result suggests that we can improve upon the previous $O(\alpha n)$ mean estimation guarantees. Furthermore, the following theorem proves that this upper bound is tight under $\mathcal{A}_3^\alpha$.

Theorem 5. *Let $\mathcal{N}(\mu, \Sigma)$ be a Gaussian with support in the range of linear transformation A and let $\mathcal{A}_3^\alpha$ adversarially corrupt the samples. Any algorithm that outputs a mean estimator $\hat{\mu}$ must satisfy $\|\mu - \hat{\mu}\|_\Sigma = \Omega(\alpha \frac{n}{m_A})$.*

While our analysis focuses on Gaussian distributions, our analysis framework generalizes to any class of distributions that admits an efficient robust mean estimator under sample-level corruption, e.g. distributions with bounded covariance. This generality stems from our general reduction scheme between coordinate-level and sample-level corruption. Such extension is also feasible for results in Section 4.

4. Efficient Algorithms

We discuss efficient estimation algorithms the match the error bounds of our analysis in Section 3.3. Building upon practical approaches in data cleaning (Wu et al., 2020; Rekatsinas et al., 2017) that leverage the structure in data to perform estimation over corrupted data, we study the practicality and theoretical guarantees obtained by such recover-and-estimate approaches for robust mean estimation.

Two-Step Meta-Algorithm The meta-algorithm has two steps: 1) *Recover*: use the dependencies across coordinates of the data (i.e., the structure) to recover the values of corrupted samples (when possible); 2) *Estimate*: After fixing corruptions, perform statistical estimation using an existing mean estimation method (e.g., empirical mean estimation). For theoretical analysis, we require *exact recovery* in the first step, i.e., we seek to recover the true sample before corruption without any errors, and view the samples that cannot be exactly recovered as the remaining corruptions.

We first show that when the corruption is limited to missing entries, there exist efficient algorithms with polynomial complexity that perform exact recovery. We show that plugging these algorithms into the two-step meta-algorithm yields practical algorithms for robust mean estimation under coordinate-level corruption, and these algorithms match the information-theoretic bounds. We further study corruption due to replacements. Here, we show that exact recovery in the presence of coordinate-level corruptions is NP-hard by building connections to sparse recovery. To overcome such hardness, we propose a randomized algorithm with a probabilistic guarantee with respect to the recovery, and as such, mean estimation.

4.1. Efficient Algorithms for Missing Entries

We show two computationally efficient algorithm instances that achieve near-optimal guarantees for known and unknown structure in the presence of missing entries. Details of the recovery step in the two algorithms are provided in the supplementary material. We assume a sufficiently large sample size (infinite in the case of unknown structure) for all the analysis.

Mean Estimation with Known Structure When matrix A is known, we recover missing coordinates as follows: we solve the linear system of equations formed by the non-corrupted data in the sample and A to estimate z, and then use this estimation to complete the missing values of x. Such a recovery step has a complexity of $O(r^3)$. Given the recovered samples, we proceed with mean estimation.

The above algorithm achieves error $\Theta(\alpha\frac{n}{m_A})$: the best strategy of $\mathcal{A}_2^\rho$ or $\mathcal{A}_3^\alpha$ is to corrupt coordinates so that recovery is impossible. To this end, a coordinate-level adversary must corrupt at least m_A coordinates for a sample to make coordinate recovery impossible. The two-step approach of recovery by solving a linear system and mean estimation with the Tukey median over full samples is information theoretically-optimal. However, the Tukey median is computationally intractable, and we use the empirical mean to obtain a computationally efficient algorithm. This approach yields a near-optimal guarantee of $\tilde{O}(\frac{\alpha n}{m_A})$ that is tight up to logarithmic factors—here $\tilde{O}(\epsilon) = O(\epsilon\sqrt{\log(1/\epsilon)})$.

Theorem 6. *Assume samples $x_i = Az_i$ and z_i comes from a Gaussian such that $x_i \sim \mathcal{N}(\mu, \Sigma)$ with support in the range of linear transformation A. Given a set of i.i.d. samples corrupted by $\mathcal{A}_2^\rho$ (when $\rho = \alpha$) or $\mathcal{A}_3^\alpha$, recover missing coordinates by solving a linear system of equations then discard all unrecoverable samples. The empirical mean $\hat{\mu}$ of the remaining samples satisfies $\|\hat{\mu} - \mu\|_\Sigma = \tilde{O}(\frac{\alpha n}{m_A})$, while the Tukey median $\hat{\mu}_{\text{Tukey}}$ of the remaining samples satisfies $\|\hat{\mu}_{\text{Tukey}} - \mu\|_\Sigma = O(\frac{\alpha n}{m_A})$.*

Theorem 6 shows that for A with $m_A \approx n$, while the strong adversary $\mathcal{A}_3^\alpha$ introduces corruptions that shift the observed distribution by $d_{\text{ENTRY}} = \alpha$, it can only affect the mean estimation as much as the weaker adversary $\mathcal{A}_1^\epsilon$ (with $\epsilon = \alpha$), which shifts the observed distribution only by $d_{\text{TV}} = \epsilon$. This result implies that recovery by leveraging the structure reduces the strength of $\mathcal{A}_3^\alpha$ (and also $\mathcal{A}_2^\rho$ with $\rho = \epsilon$) to that of $\mathcal{A}_1^\epsilon$. In fact, $m_A = n - r + 1$ for almost every A with respect to the Lebesgue measure on $\mathbb{R}^{n\times r}$, so $m_A \approx n$ when r is sufficiently low-dimensional. The above means that we can tolerate coordinate-level corruptions with large ρ and α only if we first recover and then estimate.

Mean Estimation with Unknown Structure If A is unknown, we can estimate it using the visible entries before we use them to impute the missing ones. We build on the next result: matrix completion can help robust mean estimation in the setting of $x_i = Az_i$ when A is unknown but has full rank, in which case $m_A = n - r + 1$. Corollary 1 by Pimentel-Alarcón et al. (2016) gives the conditions in which we can uniquely recover a low rank matrix with missing entries. This result goes beyond random missing values and considers deterministic missing-value patterns. We state it as follows:

Lemma 1. *Assume samples $x_i = Az_i$ and z_i comes from a Gaussian such that $x_i \sim \mathcal{N}(\mu, \Sigma)$ with support in the range of A, but A is unknown and full rank. If there exist $r + 1$ disjoint groups of $n - r$ samples, and in each group, any k samples have at least $r + k$ dimensions which are not completely hidden, all the samples with at least r visible entries can be uniquely recovered.*

The above leads to the next algorithm: We recover the missing coordinates via matrix completion. A typical algorithm

for matrix completion is iterative hard-thresholded SVD (ITHSVD) (Chunikhina et al., 2014), which has a complexity of $O(TNnr)$, where T is the maximum number of iterations and N is the sample size. Then, we use either Tukey median or a empirical mean estimation. We next analyze the guarantees of this algorithm.

Matrix completion requires learning the r-dimensional subspace spanned by the samples. Samples with more than r visible entries provide information to identify this subspace. Given the subspace, samples with at least r visible entries can be uniquely recovered. For corruptions by $\mathcal{A}_3^\alpha$ we show:

Lemma 2. *If A is unknown and the data is corrupted by $\mathcal{A}_3^\alpha$ where $\alpha \geq \frac{1}{n}$, we cannot recover any missing coordinate, otherwise we can recover all the samples with less than m_A missing entries.*

If we combine this lemma with Theorem 6, we can obtain optimal mean estimation error—we obtain the same error guarantees with Theorem 6—using matrix completion only when the budget of $\mathcal{A}_3^\alpha$ is bounded as a function of the data dimensions, i.e., when $\alpha < 1/n$ (see A.11 in the supplementary material). These guarantees are information-theoretically optimal but pessimistic: $\mathcal{A}_3^\alpha$ can hide all coordinates from the same dimension which can be unrealistic. Thus, we focus on adversary $\mathcal{A}_2^\rho$.

Theorem 7. *Assume samples $x_i = Az_i$ and z_i comes from a Gaussian such that $x_i \sim \mathcal{N}(\mu, \Sigma)$ with support in the range of A, but A is unknown and full rank. Under $\mathcal{A}_2^\rho$ where $\rho < \frac{m_A - 1}{n + (m_A - 1)(m_A - 2)}$, the above two-step algorithm obtains $\|\mu - \hat{\mu}\|_\Sigma = \tilde{O}(\frac{\rho n}{m_A})$, while the Tukey median $\hat{\mu}_{Tukey}$ of the remaining samples satisfies $\|\hat{\mu}_{Tukey} - \mu\|_\Sigma = O(\frac{\rho n}{m_A})$.*

Lemma 3. *Assume A is unknown and full rank. Under $\mathcal{A}_2^\rho$, if $\rho \geq \frac{m_A - 1}{n}$, we cannot recover any corrupted sample; if $\rho < \frac{m_A - 1}{n + (m_A - 1)(m_A - 2)}$, we can recover all samples with at least r visible entries.*

4.2. Discussion of Corruptions due to Replacements

The above two-step approach relies on exact recovery to provide guarantees for mean estimation. As we showed in the previous section, exact recovery is possible in the presence of missing values. However, when corruptions are introduced due to adverarial replacements, recovery is NP-hard in general. We show this by reducing the problem of sparse recovery to it. Sparse recovery is the problem of finding sparse solutions to underdetermined systems of linear equations, and it is shown to be computationally intractable by Berlekamp et al. (1978). Details of the connection between the two can be found in the supplementary material.

Works on decoding and signal processing have proposed efficient algorithms for exact recovery under replacements with probabilistic guarantees. Typical algorithms include basis pursuit (Candes & Tao, 2005) and orthogonal matching pursuit (Davenport & Wakin, 2010). However, these algorithms pose strict conditions on A, and have guarantees only on a limited family of matrices (e.g. Gaussian matrices). We defer the details of those conditions and guarantees to the supplementary material. For a generic A, as the matrices we consider in our problem setting, existing algorithms fail to provide any guarantee.

To alleviate the difficulty of recovery for replacements, we propose a randomized algorithm that has probabilistic guarantees for exact recovery without posing strict restrictions on matrix A. We fit this algorithm into the aforementioned two-step meta-algorithm, and provide an analysis similar to the case of missing entries for known structure.

Algorithm 1 Recovery for Coordinate-level Replacements

input: $A \in \mathbb{R}^{n \times r}$, corrupted sample $\tilde{x} \in \mathbb{R}^n$, $c > 0$
if $\exists z$ *such that* $\tilde{x} = Az$ **then**
  return $\tilde{x}$
end
for $i = 1 \dots r^c$ **do**
  Uniformly at random, select r out of n rows of A such that they are linearly independent
  $(\tilde{x}', A') \leftarrow$ linear system of the corresponding r coordinates
  Compute the solution $\hat{z}$ to $\tilde{x}' = A'z$
  Store the $\hat{z}$ that has the smallest $\|\tilde{x} - A\hat{z}\|_0$ so far
end
return $A\hat{z}$

For a corrupted sample $\tilde{x}$ that does not lie on the subspace generated by A, Algorithm 1 efficiently recovers a candidate solution $A\hat{z}$ that is not too far from $\tilde{x}$. More precisely, it returns $A\hat{z}$ such that $\|\tilde{x} - A\hat{z}\|_0 \leq \frac{r}{\ln(r)}\|\tilde{x} - x\|_0$ with high probability, assuming the original sample x is information-theoretically recoverable (less than $\frac{m_A}{2}$ coordinates are corrupted by the proof of Theorem 5). Algorithm 1 requires solving a linear system of size r multiple times, yielding a runtime of $O(nr^{4+c})$, where c is a parameter chosen depending on how accurate we want the solution to be. Since Algorithm 1 gives us a recovery routine for each corrupted sample, we have the following result on mean estimation.

Theorem 8. *Preprocessing the data with Algorithm 1 and then applying a robust mean estimator for Gaussians, e.g. of Diakonikolas et al. (2019a), yields a mean estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 = \tilde{O}(\frac{r}{\ln r} \cdot \frac{\alpha n}{m_A})$.*

Our exploration of computationally efficient algorithms for mean estimation under coordinate-level replacements are preliminary. The randomized algorithm only works when the structure is known, and the guarantee it provides is probabilistic. Finding algorithms with better guarantees and recovery methods when the structure is unknown is an exciting direction for future research.

Remark We assume sufficiently large sample size and ignore the sampling error. In fact, the sample size and the probability when the error bound holds depend on the mean estimator in the *Estimate* step of the meta algorithm. For example, in Theorem 8, if we apply the robust mean estimator for Gaussians by Diakonikolas et al. (2019a), with sample size $N = \tilde{\Omega}(\frac{m_A^2 \log^5(1/\delta)}{\alpha^2})$, the upper bound holds with probability $1 - \delta$.

5. Experiments

We compare the two-step recover-and-estimate procedure against standard robust estimators under missing entries in real-world data. We consider the following methods:

1) *Empirical Mean:* Take the mean for each coordinate, ignoring all missing entries. 2) *Data Sanitization:* Remove any samples with missing entries, and then take the mean of the rest of the data. 3) *Coordinate-wise Median (C-Median):* Take the median for each coordinate, ignoring all missing entries. 4) *Two-Step method with Matrix Completion (Two-Step-M):* Use iterative hard-thresholded SVD (Chunikhina et al., 2014) to impute missing entries and compute the mean. We use randomized SVD (Halko et al., 2011) to accelerate.

Here, we do not include the exact recovery method in Theorem 6 since the structure is unknown in real-world data. In the supplementary material, we provide synthetic experiments showing that the two-step method with exact recovery outperforms structure-agnostic methods by over 50%.

In each experiment, we inject missing values by hiding the smallest ϵ fraction of each dimension, which introduces bias to the mean. It is the worst possible $\mathcal{A}_2^\rho$ corruption for Gaussian distributions if the estimation is taking the empirical mean or median. Note that the worst $\mathcal{A}_3^\alpha$ corruption for empirical mean or median is hiding the tail of the coordinate with the highest variance. We do not include it in the experiments since such setting yields either zero error (when $\alpha < 1/n$, n is the number of coordinates) or unbounded error (when $\alpha \geq 1/n$) if we apply Two-Step-M for data with linear structure (see Lemma 2).

We show that exploiting redundancy helps improve the robustness of mean estimation. We consider five data sets with unknown structure from the UCI repository (Dua & Graff, 2017). Detailed information of these datasets is provided in the supplemental material. For all the data sets, we report the l_2 error. We use the empirical mean of the samples before corruption to approximate the true mean. We show the results in Figure 2. We find that Two-Step-M always outperforms Empirical Mean and C-Median on Breast Cancer Wisconsin, Wearable Sensor, Mice Protein Expression. For Leaf and Blood Transfusion, the error of Two-Step-M can be as much as $2\times$ lower than the error of the other methods for small budgets. The estimation error becomes high only for large values of ϵ. Data Sanitization performs worse than Empirical Mean and C-Median.

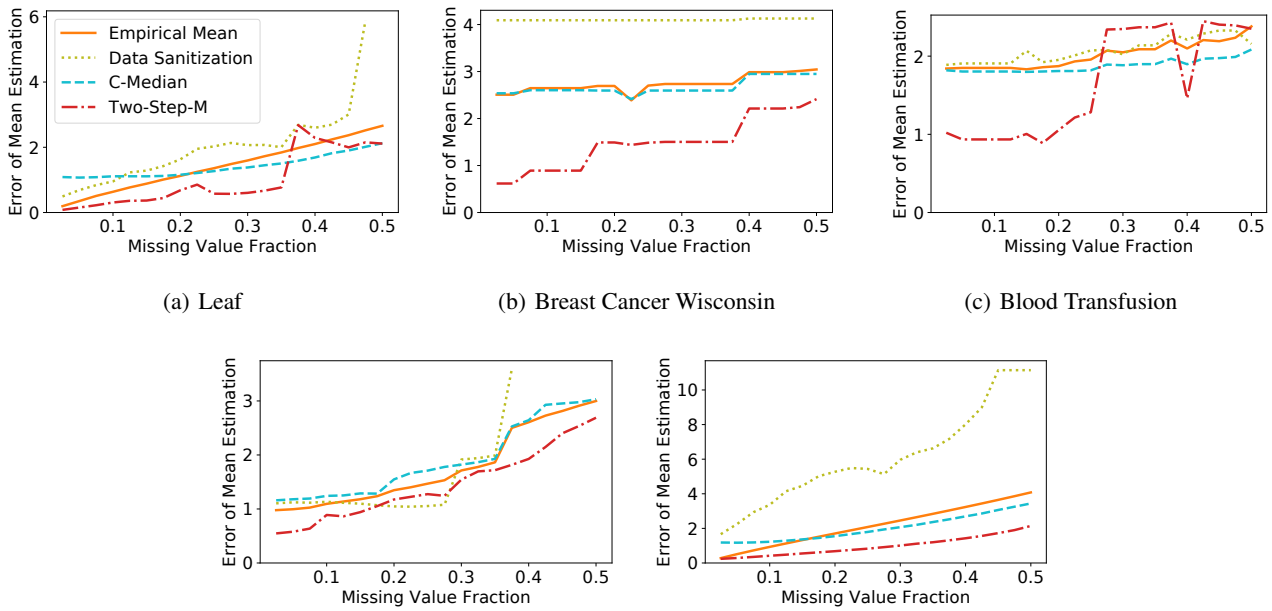
Note that the error of Two-Step-M is not always increasing as the missing value fraction increases, even for synthetic data (see the supplemental material). This is because the performance of Two-Step-M depends on both the number of samples that can be recovered and the quality of the learned structure. When the missing value fraction is low, the conditions in Lemma 1 are satisfied, and the structure is learned exactly (for synthetic data) or well approximated (for the real data) via matrix completion. In this case, the error grows monotonically. When the missing value fraction is high, the conditions in Lemma 1 cannot be satisfied, and thus the quality of the learned structure is not guaranteed. In this case, the learned structure can be accurate or not by chance, and the error may not grow monotonically.

6. Related Work

Many works have studied problems in robust statistics. These include robust mean and covariance estimation (Lai et al., 2016; Daskalakis et al., 2018; Diakonikolas et al., 2019a; Cheng et al., 2019; Kontonis et al., 2019; Zhu et al., 2019), robust optimization (Charikar et al., 2017; Diakonikolas et al., 2018; Duchi & Namkoong, 2018; Prasad et al., 2018), robust regression (Huber et al., 1973; Klivans et al., 2018; Diakonikolas et al., 2019b; Gao et al., 2020), robust subspace learning (Maunu & Lerman, 2019; Awasthi et al., 2020), and computational hardness of robustness (Hardt & Moitra, 2013; Klivans & Kothari, 2014; Diakonikolas et al., 2019b; Hopkins & Li, 2019).The works that are most relevant to ours focus on:

Entry-level Corruption Zhu et al. (2019) define a family of non-parametric distributions for which robust estimation is well-behaved under a corruption model bounded by the Wasserstein metric, while we focus on the information-theoretic analysis of mean estimation. Wang & Gu (2017) studies robust covariance estimation under a corruption model similar to $\mathcal{A}_2^\rho$. They assume sparse covariance, which is different from our low-dimensional-subspace assumption. Phocas (Xie et al., 2018) performs structure agnostic coordinate-wise estimation without considering recovery under $\mathcal{A}_2^\rho$-type corruption. Loh & Tan (2018) study learning a sparse precision matrix of the data under cell-level noise but for a d_{TV} adversary.

Data Recovery State-of-the-art methods for data imputation use data redundancy to obtain high accuracy even for systematic noise. SVD-based imputation methods (Mazumder et al., 2010; Troyanskaya et al., 2001a) assume linear relations across coordinates. There are other works that consider different models, including K-nearest neighbors, SVMs, decision trees, and even attention-based



(a) Leaf (b) Breast Cancer Wisconsin (c) Blood Transfusion

(d) Wearable Sensor (e) Mice Protein

Figure 2. Error of mean estimation on real-world data sets.

mechanisms (Bertsimas et al., 2017; Wu et al., 2020) to discover more complex non-linear structures. Our theoretical analysis provides intuition as to why these methods outperform solutions that only rely on coordinate-wise statistics.

Robust Mean Estimators Previous works have studied the problem of robust mean estimation, and proposed computationally efficient estimators for high-dimensional data, such as truncated mean (Burdett, 1996), geometric median (Cohen et al., 2016), and the iterative filtering algorithm by (Diakonikolas et al., 2016). Those works are orthogonal to ours, since they study robust mean estimation when there exhibit no structure in the data, while we focus on the effects of structure and recovery. Those estimators can be plugged into the two-step meta-algorithm, and the error bound analysis can be derived following our reduction scheme.

7. Conclusion

We studied the problem of robust mean estimation under coordinate-level corruption. We proposed d_{ENTRY}, a new measure of distribution shift for coordinate-level corruptions and introduced adversary models that capture more realistic corruptions than prior works. We presented an information-theoretic analysis of robust mean estimation for these adversaries and showed that when the data exhibits redundancy one should first fix corrupted samples before estimation. Our analysis is tight. Finally, we study the existence of practical algorithms for mean estimation that matches the information-theoretic bounds from our analysis.

Acknowledgements

The authors would like to thank Nils Palumbo for early discussions on this paper. Also, this work was supported by Amazon under an ARA Award, by NSF under grants IIS-1755676, IIS-1815538 and CCF-2008006.

References

Awasthi, P., Chatziafratis, V., Chen, X., and Vijayaraghavan, A. Adversarially robust low dimensional representations, 2020.

Berlekamp, E., McEliece, R., and van Tilborg, H. On the inherent intractability of certain coding problems (corresp.). *IEEE Transactions on Information Theory*, 24(3): 384–386, 1978. doi: 10.1109/TIT.1978.1055873.

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research*, 18(1):7133–7171, 2017.

Burdett, K. Truncated means and variances. *Economics Letters*, 52(3):263–267, 1996.

Candes, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. doi: 10.1109/TIT.2005.858979.

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

Cheng, Y., Diakonikolas, I., Ge, R., and Woodruff, D. Faster algorithms for high-dimensional robust covariance estimation. *arXiv preprint arXiv:1906.04661*, 2019.

Chunikhina, E., Raich, R., and Nguyen, T. Performance analysis for matrix completion via iterative hard-thresholded svd. In *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 392–395. IEEE, 2014.

Cohen, M. B., Lee, Y. T., Miller, G., Pachocki, J., and Sidford, A. Geometric median in nearly linear time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 9–21, 2016.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 639–649. IEEE, 2018.

Davenport, M. A. and Wakin, M. B. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Transactions on Information Theory*, 56(9):4395–4401, 2010.

Diakonikolas, I. and Kane, D. M. Recent advances in algorithmic high-dimensional robust statistics, 2019.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 655–664, 2016. doi: 10.1109/FOCS.2016.85.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 999–1008. JMLR.org, 2017.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Steinhardt, J., and Stewart, A. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

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

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

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Gao, C. et al. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Hardt, M. and Moitra, A. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pp. 354–375, 2013.

Hopkins, S. B. and Li, J. How hard is robust mean estimation? *arXiv preprint arXiv:1903.07870*, 2019.

Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pp. 492–518. Springer, 1992.

Huber, P. J. et al. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

Khosravi, P., Liang, Y., Choi, Y., and Van den Broeck, G. What to expect of classifiers? reasoning about logistic regression with missing features. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 2716–2724. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/377. URL https://doi.org/10.24963/ijcai.2019/377.

Klivans, A. and Kothari, P. Embedding hard learning problems into gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

Klivans, A., Kothari, P. K., and Meka, R. Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*, 2018.

Kontonis, V., Tzamos, C., and Zampetakis, M. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1578–1595. IEEE, 2019.

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

Loh, P.-L. and Tan, X. L. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electron. J. Statist.*, 12(1):1429–1467, 2018. doi: 10.1214/18-EJS1427. URL https://doi.org/10.1214/18-EJS1427.

Maunu, T. and Lerman, G. Robust subspace recovery with adversarial outliers, 2019.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug): 2287–2322, 2010.

Pimentel-Alarcón, D. L., Boston, N., and Nowak, R. D. A characterization of deterministic sampling patterns for low-rank matrix completion. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):623–636, 2016.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Rekatsinas, T., Chu, X., Ilyas, I. F., and Ré, C. Holoclean: Holistic data repairs with probabilistic inference. *PVLDB*, 10(11):1190–1201, 2017. doi: 10.14778/3137628.3137631. URL http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001a.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001b.

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

Wang, L. and Gu, Q. Robust gaussian graphical model estimation with arbitrary corruption. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3617–3626. JMLR. org, 2017.

Wu, R., Zhang, A., Ilyas, I., and Rekatsinas, T. AimNet: attention-based learning for missing data imputation. MLsys'20, 2020.

Xie, C., Koyejo, O., and Gupta, I. Phocas: dimensional byzantine-resilient stochastic gradient descent, 2018.

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