2016 IEEE 57th Annual Symposium on Foundations of Computer Science

Robust Estimators in High Dimensions without the Computational Intractability

Ilias Diakonikolas
CS
University of Southern California
Los Angeles, CA, USA
diakonik@usc.edu

Gautam Kamath
EECS
Massachusetts Institute of Technology
Cambridge, MA, USA
g@csail.mit.edu

Daniel M. Kane
CSE & Math
University of California San Diego
La Jolla, CA, USA
dakane@cs.ucsd.edu

Jerry Li
EECS
Massachusetts Institute of Technology
Cambridge, MA, USA
jerryzli@mit.edu

Ankur Moitra
Math
Massachusetts Institute of Technology
Cambridge, MA, USA
moitra@mit.edu

Alistair Stewart
CS
University of Southern California
Los Angeles, CA, USA
alistais@usc.edu

***Abstract*—We study high-dimensional distribution learning in an agnostic setting where an adversary is allowed to arbitrarily corrupt an epsilon fraction of the samples. Such questions have a rich history spanning statistics, machine learning and theoretical computer science. Even in the most basic settings, the only known approaches are either computationally inefficient or lose dimension dependent factors in their error guarantees. This raises the following question: Is high-dimensional agnostic distribution learning even possible, algorithmically?**

In this work, we obtain the first computationally efficient algorithms for agnostically learning several fundamental classes of high-dimensional distributions: (1) a single Gaussian, (2) a product distribution on the hypercube, (3) mixtures of two product distributions (under a natural balancedness condition), and (4) mixtures of k Gaussians with identical spherical covariances. All our algorithms achieve error that is independent of the dimension, and in many cases depends nearly-linearly on the fraction of adversarially corrupted samples. Moreover, we develop a general recipe for detecting and correcting corruptions in high-dimensions, that may be applicable to many other problems.

***Keywords*-unsupervised learning, statistical learning, density estimation robust algorithm**

I. Introduction

A. Background

A central goal of machine learning is to design efficient algorithms for fitting a model to a collection of observations. In recent years, there has been considerable progress on a variety of problems in this domain, including algorithms with provable guarantees for learning mixture models [1], [2], [3], [4], [5], phylogenetic trees [6], [7], HMMs [8], topic models [9], [10], and independent component analysis [11]. These algorithms crucially rely on the assumption that the observations were actually generated by a model in the family. However, this simplifying assumption is not meant to be exactly true, and it is an important direction to explore what happens when it holds only in an approximate sense. In this work, we study the following family of questions:

Question I.1. *Let $\mathcal{D}$ be a family of distributions on $\mathbb{R}^d$. Suppose we are given samples generated from the following process: First, m samples are drawn from some unknown distribution P in $\mathcal{D}$. Then, an adversary is allowed to arbitrarily corrupt an ε-fraction of the samples. Can we efficiently find a distribution P' in $\mathcal{D}$ that is $f(\varepsilon, d)$-close, in total variation distance, to P?*

This is a natural formalization of the problem of designing robust and efficient algorithms for distribution estimation. We refer to it as *(proper) agnostic distribution learning* and we refer to the samples as being ε-corrupted. This family of problems has its roots in many fields, including statistics, machine learning, and theoretical computer science. Within computational learning theory, it is related to the agnostic learning model of Haussler [12] and Kearns *et al.* [13], where the goal is to learn a labeling function whose agreement with some underlying target function is close to the best possible, among all functions in some given class. In the even more challenging malicious noise model [14], [15], an adversary is allowed to corrupt both the labels and the samples. A major difference with our setting is that these models apply to supervised learning problems, while here we will work in an unsupervised setting.

Within statistics and machine learning, inference problems like Question I.1 are often termed "estimation under model misspecification". The usual prescription is to use the maximum likelihood estimator [16], [17], which is unfortunately hard to compute in general. Even

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.85

ignoring computational considerations, the maximum likelihood estimator is only guaranteed to converge to the distribution P' in $\mathcal{D}$ that is closest (in Kullback-Leibler divergence) to the distribution from which the observations are generated. This is problematic because such a distribution is not necessarily close to P at all.

A branch of statistics – called robust statistics [18], [19] – aims to tackle questions like the one above. The usual formalization is in terms of breakdown point, which (informally) is the fraction of observations that an adversary would need to control to be able to completely corrupt an estimator. In low-dimensions, this leads to the prescription that one should use the empirical median instead of the empirical mean to robustly estimate the mean of a distribution, and interquartile range for robust estimates of the variance. In high-dimensions, the Tukey depth [20] is a high-dimensional analogue of the median that, although provably robust, is hard to compute [21]. Similar hardness results have been shown [22], [23] for essentially all known estimators in robust statistics.

Is high-dimensional agnostic distribution learning even possible, algorithmically? The difficulty is that corruptions are often hard to detect in high dimensions, and could bias the natural estimator by dimension-dependent factors. In this work, we study agnostic distribution learning for a number of fundamental classes of distributions: (1) a single Gaussian, (2) a product distribution on the hypercube, (3) mixtures of two product distributions (under a natural balancedness condition), and (4) mixtures of k Gaussians with identical spherical covariances. Prior to our work, all known efficient algorithms (e.g. [24], [25]) for these classes required the error guarantee, $f(\varepsilon, d)$, to depend *polynomially* in the dimension d. Hence, previous efficient estimators could only tolerate at most a $1/\operatorname{poly}(d)$ fraction of errors. In this work, we obtain *the first* efficient algorithms for the aforementioned problems, where $f(\varepsilon, d)$ is *completely independent of* d and depends polynomially (often, nearly linearly) in the fraction ε of corrupted samples. Our work is just a first step in this direction, and there are many exciting questions left to explore.

B. Our Techniques

All of our algorithms are based on a common recipe. The first question to address is the following: Even if we were given a candidate hypothesis P', how could we test if it is ε-close in total variation distance to P? The usual way to certify closeness is to exhibit a coupling between P and P' that marginally samples from both distributions, where the samples produced from each agree with probability $1 - \varepsilon$. However, we have no control over the process by which samples are generated from P, in order to produce such a coupling. And even then, the way that an adversary decides to corrupt samples can introduce complex statistical dependencies.

We circumvent this issue by working with an appropriate notion of parameter distance, which we use as a proxy for the total variation distance between two distributions in the class $\mathcal{D}$. Various notions of parameter distance underly several efficient algorithms for distribution learning in the following sense. If θ and θ' are two sets of parameters that define distributions P_θ and $P_{\theta'}$ in a given class $\mathcal{D}$, a learning algorithm often relies on establishing the following type of relation[1] between $d_{\mathrm{TV}}(P_\theta, P_{\theta'})$ and the parameter distance $d_p(\theta, \theta')$:

$$\operatorname{poly}(d_p(\theta,\theta'), 1/d) \le d_{\mathrm{TV}}(P_\theta, P_{\theta'}) \le \operatorname{poly}(d_p(\theta,\theta'), d)\ . \tag{1}$$

Unfortunately, in our agnostic setting, we cannot afford for (1) to depend on the dimension d at all. Any such dependence would appear in the error guarantee of our algorithm. Instead, the starting point of our algorithms is a notion of parameter distance that satisfies

$$\operatorname{poly}(d_p(\theta,\theta')) \le d_{\mathrm{TV}}(P_\theta, P_{\theta'}) \le \operatorname{poly}(d_p(\theta,\theta')) \tag{2}$$

which allows us to reformulate our goal of designing robust estimators, with distribution-independent error guarantees, as the goal of robustly estimating θ according to d_p. In several settings, the choice of the parameter distance is rather straightforward. It is often the case that some variant of the ℓ_2-distance between the parameters works[2].

Given our notion of parameter distance satisfying (2), our main ingredient is an efficient method for robustly estimating the parameters. We provide two algorithmic approaches which are based on similar principles. Our first approach is faster, requiring only approximate eigenvalue computations. Our second approach relies on convex programming and achieves much better sample complexity, in some cases matching the information-theoretic limit. Notably, either approach can be used to give all of our concrete learning applications with nearly identical error guarantees. In what follows, we specialize to the problem of robustly learning the mean μ of a Gaussian whose covariance is promised to be the identity, which we will use to illustrate how both approaches operate. We emphasize that what is needed

[1] For example, the work of Kalai *et al.* [2] can be reformulated as showing that for any pair of mixtures of two Gaussians (with suitably bounded parameters), the following quantities are polynomially related: (1) discrepancy in their low-order moments, (2) their parameter distance, and (3) their total variation distance. This ensures that any candidate set of parameters that produce almost identical moments must itself result in a distribution that is close in total variation distance.

[2] This discussion already points to why it may be challenging to design agnostic algorithms for mixtures of arbitrary Gaussians or arbitrary product distributions: It is not clear what notion of parameter distance is polynomially related to the total variation distance between two such mixtures, without any dependence on d.

to learn the parameters in more general settings requires many additional ideas.

Our first algorithmic approach is an iterative greedy method that, in each iteration, filters out some of the corrupted samples. Given a set of samples S' that contains a set S of uncorrupted samples, an iteration of our algorithm either returns the sample mean of S' or finds a *filter* that allows us to efficiently compute a set $S'' \subset S'$ that is much closer to S. Note the sample mean $\widehat{\mu} = \sum_{i=1}^{N} (1/N) X_i$ (even after we remove points that are obviously outliers) can be $\Omega(\varepsilon\sqrt{d})$-far from the true mean in ℓ_2-distance. The filter approach shows that either the sample mean is already a good estimate for μ or else there is an elementary spectral test that rejects some of the corrupted points and almost none of the uncorrupted ones. The crucial observation is that if a small number of corrupted points are responsible for a large change in the sample mean, it must be the case that many of the error points are very far from the mean in some particular direction. Thus, we obtain our filter by computing the top absolute eigenvalue of a modified sample covariance matrix.

Our second algorithmic approach relies on convex programming. Here, instead of rejecting corrupted samples, we compute appropriate *weights* w_i for the samples X_i, so that the weighted empirical average $\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i$ is close to μ. We work with the convex set:

$$\mathcal{C}_\delta = \Bigg\{ w_i : 0 \le w_i \le \frac{1}{(1-\varepsilon)N}, \sum_{i=1}^{N} w_i = 1, \\ \left\| \sum_{i=1}^{N} w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \le \delta \Bigg\} .$$

We prove that *any* set of weights in $\mathcal{C}_\delta$ yields a good estimate $\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i$ in the obvious way. The catch is that the set $\mathcal{C}_\delta$ is defined based on μ, *which is unknown*. Nevertheless, it turns out that we can use the same types of spectral arguments that underly the filtering approach to design an approximate separation oracle for $\mathcal{C}_\delta$. Combined with standard results in convex optimization, this yields an algorithm for robustly estimating μ.

The third and final ingredient are some new concentration bounds. In both of the approaches above, at best we are hoping that we can remove all of the corrupted points and be left with only the uncorrupted ones, and then use standard estimators (e.g., the empirical average) on them. However, an adversary could have removed an ε-fraction of the samples in a way that biases the empirical average of the remaining uncorrupted samples. What we need are concentration bounds that show for sufficiently large N, for samples $X_1, X_2, \ldots, X_N$ from a Gaussian with mean μ and identity covariance, that every $(1-\varepsilon)N$ set of samples produces a good estimate for μ. In some cases, we can derive such concentration bounds by appealing to known concentration inequalities and taking a union bound. However, in other cases (e.g., concentration bounds for degree two polynomials of Gaussian random variables) the existing concentration bounds are not strong enough, and we need other arguments to prove that every set of $(1-\varepsilon)N$ samples produces a good estimate. Also in Section VIII we explain why some other natural strategies for robust distribution learning obtain poor guarantees in high-dimensions.

C. Our Results

We give the first efficient algorithms for agnostically learning several important distribution classes with dimension-independent error guarantees. Our first main result is for a single arbitrary Gaussian with mean μ and covariance Σ, which we denote by $\mathcal{N}(\mu, \Sigma)$. In the previous subsection, we described our convex programming approach for learning the mean vector when the covariance is promised to be the identity. A technically more involved version of the technique can handle the case of zero mean and unknown covariance. More specifically, consider the following convex set, where Σ is the unknown covariance matrix:

$$\mathcal{C}_\delta = \Bigg\{ w_i : 0 \le w_i \le \frac{1}{(1-\varepsilon)N}, \sum_{i=1}^{N} w_i = 1, \\ \left\| \Sigma^{-1/2} \left(\sum_{i=1}^{N} w_i X_i X_i^T \right) \Sigma^{-1/2} - I \right\|_F \le \delta \Bigg\} .$$

We design an approximate separation oracle for this unknown convex set, by analyzing the spectral properties of the fourth moment tensor of a Gaussian. Combining these two intermediate results, we obtain our first main result (below). Throughout this paper, we will abuse notation and write $N \ge \widetilde{\Omega}(f(d, \varepsilon, \tau))$ when referring to our sample complexity, to signify that our algorithm works if $N \ge C f(d, \varepsilon, \tau) \mathrm{polylog}(f(d, \varepsilon, \tau))$ for a large enough universal constant C.

Theorem I.2. *Let μ, Σ be arbitrary and unknown, and let $\varepsilon, \tau > 0$. There is a polynomial time algorithm which given ε, τ, and an ε-corrupted set of N samples from $\mathcal{N}(\mu, \Sigma)$ with $N \ge \widetilde{\Omega}\left(\frac{d^3 \log^2(1/\tau)}{\varepsilon^2}\right)$, produces $\widehat{\mu}$ and $\widehat{\Sigma}$ so that with probability $1-\tau$ we have $d_{\mathrm{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})) \le O(\varepsilon \log^{3/2}(1/\varepsilon))$.*

We can alternatively establish Theorem I.2 with a slightly worse sample complexity via our filtering technique. We defer the details to the full version.

Our second agnostic learning result is for a product distribution on the hypercube – arguably the most fundamental discrete high-dimensional distribution. We solve this problem using our filter technique, though our convex programming approach would also yield similar results. We start by analyzing the balanced case, when no coordinate is very close to being deterministic. This special case is interesting in its own right and captures the essential ideas of our more involved analysis for the general case. The reason is that, for two balanced product distributions, the ℓ_2-distance between their means is equivalent to their total variation distance (up to a constant factor). This leads to a clean and elegant presentation of our spectral arguments. For an arbitrary product distribution, we handle the coordinates that are essentially deterministic separately. Moreover, we use the χ^2-distance between the means as the parameter distance and, as a consequence, we need to apply the appropriate corrections to the covariance matrix. Formally, we prove:

Theorem I.3. *Let Π be an unknown binary product distribution, and let $\varepsilon, \tau > 0$. There is a polynomial time algorithm which given ε, τ, and an ε-corrupted set of N samples from Π with $N \geq \Omega\left(\frac{d^6 \log(1/\tau)}{\varepsilon^3}\right)$, produces a binary product distribution $\widetilde{\Pi}$ so that with probability $1-\tau$, we have $d_{\mathrm{TV}}(\Pi, \widetilde{\Pi}) \leq O(\sqrt{\varepsilon \log(1/\varepsilon)})$.*

For the sake of simplicity in the presentation, we did not make an effort to optimize the sample complexity of our robust estimators. We also remark that for the case of balanced binary product distributions, our algorithm achieves an error of $O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

Interestingly enough, the above two distribution classes are trivial to learn in the noiseless case, but in the agnostic setting the learning problem turns out to be surprisingly challenging. Using additional ideas, we are able to generalize our agnostic learning algorithms to *mixtures* of the above classes under some natural conditions. We note that learning mixtures of the above families is rather non-trivial even in the noiseless case. First, we study 2-mixtures of c-balanced products, which stipulates that the coordinates of the mean vector of each component are in the range $(c, 1-c)$. We prove:

Theorem I.4 (informal). *Let Π be an unknown mixture of two c-balanced binary product distribution, and let $\varepsilon, \tau > 0$. There is a polynomial time algorithm which given ε, τ, and an ε-corrupted set of N samples from Π with $N \geq \Omega\left(\frac{d^4 \log(1/\tau)}{\varepsilon^{13/6}}\right)$, produces a mixture of two binary product distributions $\widetilde{\Pi}$ so that with probability $1-\tau$, we have $d_{\mathrm{TV}}(\Pi, \widetilde{\Pi}) \leq O_c(\varepsilon^{1/6})$.*

This generalizes the algorithm of Freund and Mansour [33] to the agnostic setting. An interesting open question is to improve the ε-dependence in the above bound to (nearly) linear, or to remove the assumption of balancedness and obtain an agnostic algorithm for mixtures of two arbitrary product distributions.

Finally, we give an agnostic learning algorithm for mixtures of spherical Gaussians.

Theorem I.5. *Let $\mathcal{M}$ be a mixture of k Gaussians with spherical covariances, and let $\varepsilon, \tau > 0$ and k be a constant. There is a polynomial time (for constant k) algorithm which given ε, τ, and an ε-corrupted set of N samples from $\mathcal{M}$ with $N \geq \mathrm{poly}(k, d, 1/\varepsilon, \log(1/\tau))$, outputs an $\mathcal{M}'$ so that with probability $1-\tau$, we have $d_{\mathrm{TV}}(\mathcal{M}, \mathcal{M}') \leq \tilde{O}(\mathrm{poly}(k) \cdot \sqrt{\varepsilon})$.*

Our agnostic algorithms for (mixtures of) balanced product distributions and for (mixtures of) spherical Gaussians are conceptually related, since in both cases the goal is to robustly learn the means of each component with respect to ℓ_2-distance.

In total, these results give new robust and computationally efficient estimators for several well-studied distribution learning problems that can tolerate a constant fraction of errors independent of the dimension. This points to an interesting new direction of making robust statistics algorithmic. The general recipe we have developed here gives us reason to be optimistic about many other problems in this domain.

D. Discussion and Related Work

Our results fit in the framework of *density estimation*, a classical problem in statistics with a rich history and extensive literature (see e.g., [27], [28], [29], [30], [31]). During the past couple of decades, a body of work in theoretical computer science has been studying these questions from a computational complexity perspective; see e.g., [32], [33], [34], [35], [36], [2], [3], [37], [38], [39], [40], [41], [42], [43]. Efficient agnostic learning algorithms have been given for various one-dimensional distribution classes, but very little is known in the high-dimensional setting that we study here.

Question I.1 also resembles learning in the presence of malicious errors [14], [15]. There, an algorithm is given samples from a distribution along with their labels according to an unknown target function. The adversary is allowed to corrupt an ε-fraction of both the samples and their labels. A sequence of works studied this problem for the class of halfspaces [44], [45], [46] in the setting where the underlying distribution is a Gaussian, culminating in the work of Awasthi *et al.* [47], who gave an efficient algorithm that finds a halfspace with agreement $O(\varepsilon)$. Our work and theirs are not directly comparable, since we work in an unsupervised setting. Moreover, their algorithms need to assume that the underlying Gaussian distribution is in isotropic position. In fact, our results are complementary to theirs: One could use our algorithms (on the unlabeled examples) to learn an affine transformation that puts the underlying Gaussian distribution in approximately isotropic position, even in the presence of malicious errors, so that one can then directly apply the [47] algorithm.

Another connection is to the work on robust principal component analysis (PCA). PCA is a transformation that (among other things) is often justified as being able to find the affine transformation $Y = \Sigma^{-1/2}(X - \mu)$ that would place a collection of Gaussian random variables in isotropic position. One can think of our results on agnostically learning a Gaussian as a type of robust PCA that tolerates gross corruptions, where entire samples are corrupted. This is different than other variants of the problem where random sets of coordinates of the points are corrupted [48], or where the uncorrupted points were assumed to lie in a low-dimensional subspace to begin with [49], [50]. Finally, Brubaker [51] studied the problem of clustering samples from a *well-separated* mixture of Gaussians in the presence of adversarial noise. The goal of [51] was to separate the Gaussian components from each other, while the adversarial points are allowed to end up in any of clusters. Our work is orthogonal to [51], since even if such a clustering is given, the problem still remains to estimate the parameters of each component.

E. Comparison with [52]

In concurrent and independent work, Lai, Rao and Vempala [52] also study high-dimensional agnostic learning. Their results work for more general types of distributions, but our guarantees are stronger when learning a Gaussian. Our results are similar when the mean is unknown and the covariance is promised to be the identity. But when the covariance is also unknown, their algorithm estimates the mean and covariance to within error $O(\sqrt{\varepsilon\|\Sigma\|_2 \log d})$ and $O(\sqrt{\varepsilon \log d}\|\Sigma\|_2)$, measured in ℓ_2 norm and Frobenius norm respectively. However, such guarantees do not directly imply bounds on the total variation distance (which is our main focus), because one needs to estimate the parameters with respect to Mahalanobis distance. In contrast, by virtue of being close in total variation distance, our estimates for the mean and covariance are within $\tilde{O}(\varepsilon\sqrt{\|\Sigma\|_2})$ and $\tilde{O}(\varepsilon\|\Sigma\|_2)$ of the true values, again measured in ℓ_2 norm and Frobenius norm respectively. An interesting open question is to bridge these two works – what are the most general families of distributions for which one can obtain nearly optimal agnostic learning guarantees?

II. Preliminaries

In this section, we will introduce some basic terminology that we will use throughout. Recall that our adversary is allowed to observe N samples $X_1, X_2, \ldots X_N$ and then corrupt an ε-fraction of them arbitrarily.

Definition II.1. *Let $G \subseteq [N]$ denote the indices of the uncorrupted samples, and we let $E \subseteq [N]$ denote the indices of the corrupted samples.*

An important algorithmic object for us will be the following set, which is designed to capture the notion of selecting a set of $(1-\varepsilon)N$ samples from N samples:

Definition II.2. *For any $\frac{1}{2} > \varepsilon > 0$ and any integer N, let*

$$S_{N,\varepsilon} = \left\{ w_i : \sum_{i=1}^{N} w_i = 1, 0 \le w_i \le \frac{1}{(1-\varepsilon)N}, \forall i \right\} .$$

Given $w \in S_{N,\varepsilon}$ we will use the following notation

$$w_g = \sum_{i \in G} w_i \text{ and } w_b = \sum_{i \in E} w_i$$

to denote the total weight on good and bad points respectively. The following fact is immediate from $|E| \le \varepsilon N$ and the properties of $S_{N,\varepsilon}$.

Fact II.3. *If $w \in S_{N,\varepsilon}$ and $|E| \le \varepsilon N$, then $w_b \le \frac{2\varepsilon}{1-\varepsilon}$. Moreover, the renormalized weights w' on good points given by $w'_i = \frac{w_i}{w_g}$ for all $i \in G$, and $w'_i = 0$ otherwise, satisfy $w' \in S_{N,4\varepsilon}$, provided that $\varepsilon \le 1/6$.*

III. A Gaussian with Unknown Mean

In this section, we consider the problem of approximating μ given N samples from $\mathcal{N}(\mu, I)$ in the full adversary model. Recall that our algorithm is based on the following convex set.

$$\mathcal{C}_\delta = \left\{ w \in S_{N,\varepsilon} : \left\| \sum_{i=1}^{N} w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \le \delta \right\} .$$

It is not hard to show that $\mathcal{C}_\delta$ is non-empty for reasonable values of δ (and we will show this later). Moreover we will show that for any set of weights w in $\mathcal{C}_\delta$, the empirical average $\widehat{\mu} = \sum_{i=1}^{N} w_i X_i$ will be a good estimate for μ. The challenge is that since μ itself is unknown, there is not an obvious way to design a separation oracle for $\mathcal{C}_\delta$ even though it is convex. Our algorithm will run in two basic steps. First, it will run a very naive outlier detection to remove any points which are more than $O(\sqrt{d})$ away from the good points. These points are sufficiently far away that a very basic test can detect them. Then, with the remaining points, it will use the approximate separation oracle given below to approximately optimize with respect to C_δ. It will then take the outputted set of weights and output the empirical mean with these weights. We will explain these steps in detail below.

Our results will hold under the following deterministic conditions:

$$\left\| \sum_{i\in G} w_i(X_i-\mu)(X_i-\mu)^T - w_g I \right\|_2 \le \delta_1, \quad \text{for all } w \in S_{N,4\varepsilon}, \text{ and} \tag{3}$$

$$\left\| \sum_{i\in G} w_i(X_i-\mu) \right\|_2 \le \delta_2, \quad \text{for all } w \in S_{N,4\varepsilon} \ . \tag{4}$$

The first step in our analysis is elementary, and we apply a naive pruning strategy (based on considering pairwise distances among the points) to ensure that for all the remaining points we have $\|X_i-\mu\| \le O\left(\sqrt{d\log(N/\tau)}\right)$. From this point on, we will assume that this has already been done and we defer the description of the algorithm and its analysis to the full version. In the full version, we give concentration bounds that show that each of these conditions holds with probability at least $1-\tau$ for $\delta_1, \delta_2 = O(\varepsilon\sqrt{\log 1/\varepsilon})$ and $N = O\left(\frac{d+\log(1/\tau)}{\min(\delta_1,\delta_2)^2}\right)$.

Instead, we focus on how to design an approximate separation oracle for $\mathcal{C}_\delta$ which is our main result in the section. We will also require the following elementary bound:

Fact III.1. *Let $\mu_1, \mu_2 \in \mathbb{R}^d$ be arbitrary. Then $d_{\mathrm{TV}}(\mathcal{N}(\mu_1, I), \mathcal{N}(\mu_2, I)) \le \frac{1}{2}\|\mu_2 - \mu_1\|_2$.*

Our first step is to show that any set of weights that does not yield a good estimate for μ cannot be in the set $\mathcal{C}_\delta$:

Lemma III.2. *Suppose that (3)-(4) holds. Let $\delta = \max(\delta_1, \delta_2)$. Let $w \in S_{N,\varepsilon}$ and set $\widehat{\mu} = \sum_{i=1}^N w_i X_i$ and $\Delta = \mu - \widehat{\mu}$. Further, suppose that $\|\Delta\|_2 \ge \Omega(\delta)$. Then*

$$\left\| \sum_{i=1}^N w_i(X_i-\mu)(X_i-\mu)^T - I \right\|_2 \ge \Omega\left(\frac{\|\Delta\|^2}{\varepsilon}\right) .$$

We defer the proof to the full version. As an immediate corollary, we find that *any* set of weights in $\mathcal{C}_\delta$ immediately yields a good estimate for μ:

Corollary III.3. *Suppose that (3) and (4) hold. Let $w \in \mathcal{C}_\delta$ for $\delta = O(\varepsilon \log 1/\varepsilon)$. Then*

$$\|\Delta\|_2 \le O(\varepsilon\sqrt{\log 1/\varepsilon}) .$$

Our main result in this section is an approximate separation oracle for $\mathcal{C}_\delta$ with $\delta = O(\varepsilon \log 1/\varepsilon)$.

Theorem III.4. *Fix $\varepsilon > 0$, and let $\delta = O(\varepsilon \log 1/\varepsilon)$. Suppose that (3) and (4) hold. Let w^* denote the weights which are uniform on the uncorrupted points. Then, there is a constant c and an algorithm so that:*

1) *(Completeness) If $w = w^*$, then it outputs "YES".*
2) *(Soundness) If $w \notin \mathcal{C}_{c\delta}$, the algorithm outputs a hyperplane $\ell : \mathbb{R}^N \to \mathbb{R}$ so that $\ell(w) \ge 0$ but $\ell(w^*) < 0$.*

These two facts imply that the ellipsoid method with this separation oracle will terminate in $\mathrm{poly}(d, 1/\varepsilon)$ steps, and moreover with high probability output a w' so that $\|w - w'\|_\infty < \varepsilon/(N\sqrt{d\log(N/\tau)})$, for some $w \in \mathcal{C}_{c\delta}$. Moreover, it will do so in polynomially many iterations.

The separation oracle is given in Algorithm 1.

Algorithm 1 Separation oracle sub-procedure for agnostically learning the mean.

```
function SepOracleUnknownMean($w$)
    Let $\widehat{\mu} = \sum_{i=1}^N w_i X_i$.
    For $i = 1, \ldots, N$, define $Y_i = X_i - \widehat{\mu}$.
    Let $\lambda$ be the eigenvalue of largest magnitude of $M = \sum_{i=1}^N w_i Y_i Y_i^T - I$.
    Let $v$ be its associated eigenvector.
    if $|\lambda| < \frac{c}{2}\delta$ then
        return "YES".
    else if $\lambda > \frac{c}{2}\delta$ then
        return the hyperplane
            $\ell(w) = \left(\sum_{i=1}^N w_i \langle Y_i, v\rangle^2 - 1\right) - \lambda.$
    else
        return the hyperplane
            $\ell(w) = \lambda - \left(\sum_{i=1}^N w_i \langle Y_i, v\rangle^2 - 1\right).$
```

Next, we prove correctness for our approximate separation oracle:

Proof of Theorem III.4: Again, let $\Delta = \mu - \widehat{\mu}$. By expanding out the formula for M, we get:

$$\begin{aligned} &\sum_{i=1}^N w_i Y_i Y_i^T - I \\ &= \sum_{i=1}^N w_i (X_i-\mu)(X_i-\mu)^T - I - \Delta\Delta^T . \end{aligned}$$

Let us now prove completeness.

Claim III.5. *Suppose $w = w^*$. Then $\|M\|_2 < \frac{c}{2}\delta$.*

Proof: Recall that w^* are the weights that are uniform on the uncorrupted points. Because $|E| \le 2\varepsilon N$, we have that $w^* \in S_{N,\varepsilon}$. We can now use (3) to conclude that $w^* \in \mathcal{C}_{\delta_1}$. Now, by Corollary III.3 we have that $\|\Delta\|_2 \leq O(\varepsilon\sqrt{\log 1/\varepsilon})$. Thus,

$$\begin{aligned}&\left\|\sum_{i=1}^N w_i^*(X_i-\mu)(X_i-\mu)^T - I - \Delta\Delta^T\right\|_2\\ &\leq \left\|\sum_{i=1}^N w_i^*(X_i-\mu)(X_i-\mu)^T - I\right\|_2 + \|\Delta\Delta^T\|_2\\ &< \frac{c\delta}{2}\ .\end{aligned}$$

■

We now turn our attention to soundness.

Claim III.6. *Suppose that $w \notin C_{c\delta}$. Then $|\lambda| > \frac{c}{2}\delta$.*

Proof: By the triangle inequality, we have

$$\begin{aligned}&\left\|\sum_{i=1}^N w_i(X_i-\mu)(X_i-\mu)^T - I - \Delta\Delta^T\right\|_2\\ &\geq \left\|\sum_{i=1}^N w_i(X_i-\mu)(X_i-\mu)^T - I\right\|_2 - \left\|\Delta\Delta^T\right\|_2\ .\end{aligned}$$

Let us now split into two cases. If $\|\Delta\|_2 \leq \sqrt{c\delta/10}$, then the first term above is at least $c\delta$ by definition and we can conclude that $|\lambda| > c\delta/2$. On the other hand, if $\|\Delta\|_2 \geq \sqrt{c\delta/10}$, by Lemma III.2, we have that

$$\begin{aligned}&\left\|\sum_{i=1}^N w_i(X_i-\mu)(X_i-\mu)^T - I - \Delta\Delta^T\right\|_2\\ &\geq \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right) - \|\Delta\|_2^2 = \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right) \qquad (5)\end{aligned}$$

which for sufficiently small ε also yields $|\lambda| > c\delta/2$. ■

Now by construction $\ell(w) \geq 0$. The last step is to establish the following claim:

Claim III.7. $\ell(w^*) < 0$.

The proof involves a case analysis on Δ, and follows by elementary manipulations. We defer the proof to the full version. ■

IV. Mixtures of Spherical Gaussians

In the full version of our paper, we give algorithms for learning mixtures of k spherical Gaussians. For ease of exposition, in this extended abstract we focus on the case where all component covariances are identical to the identity. The main idea is that the techniques developed in [53] for learning such mixtures only require us to learn a sufficiently good estimate of the true covariance, which is given by $I + \sum_{j=1}^k \alpha_j(\mu_j - \mu)(\mu_j - \mu)^T$, where $\mu = \mathbb{E}_{X\sim F}[X]$, where α_j and μ_j are the mixing weights and means of each component. In contrast to our approach above, we do not know the covariance of the mixture. However we still have the useful property that after subtracting off I the resulting covariance (without any corruptions) is low rank. Thus, in the definition of C_δ, instead of insisting that the error has low spectral norm, we insist the sum of the top k eigenvalues of the error cannot be too large. By similar but somewhat more involved calculations as in the unknown mean case, this allows us to either cluster the components, or recover an estimate of the covariance up to spectral error $\widetilde{O}(\varepsilon)$. We can then directly appeal to the techniques in [53] which allows us to learn the k-GMM up to error $\widetilde{O}(\sqrt{\varepsilon})$.

V. A Gaussian with Unknown Covariance

In this section we study the problem of agnostically learning a Gaussian with zero mean and unknown covariance Σ. Our result for agnostically learning a single Gaussian, where both the mean and covariance are unknown then follows in a straightforward manner by combining the algorithm in this section and the one in Section III, but we defer the details to the full version of our paper. We require the following bound:

Fact V.1. *Let $\Sigma_1, \Sigma_2 \succ 0$. Then*

$$\begin{aligned}&d_{\mathrm{TV}}(\mathcal{N}(0,\Sigma_1), \mathcal{N}(0,\Sigma_2))\\ &\leq O\left(\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I\|_F\right)\ .\end{aligned}$$

Recall, our algorithm is based on the following convex set:

$$\begin{aligned}C_\delta = &\{w \in S_{N,\varepsilon} :\\ &\left.\left\|\Sigma^{-1/2}\left(\sum_{i=1}^m w_i X_i X_i^T\right)\Sigma^{-1/2} - I\right\|_F \leq \delta\right\}\ .\end{aligned}$$

Again, we design an approximate separation oracle for this set. We do so by exploiting the fact that if $w \notin C_{c\delta}$, the corrupted points must contribute disproportionately in some way that we can detect, spectrally. We use second order statistics of the covariance — namely the fourth moment tensor. We establish a number of new concentration bounds for the empirical fourth moment tensor in order to analyze our algorithm, which may be of independent interest. Apart from this, the main technical difficulty is that we do not know the exact form of the fourth moment tensor because it depends on Σ. It turns out that considering a restricted eigenvalue problem on a carefully designed subspace, we can show that the contribution of the corrupted points is the dominant term. We are then able to compute an estimate $\widehat{\Sigma}$ with $\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I\|_F \leq \widetilde{O}(\varepsilon)$ which by Fact V.1 gives an estimate which is $\widetilde{O}(\varepsilon)$-close in total variation distance. We defer the details to the full version of our paper.

VI. Binary Product Distributions

In this section, we study the problem of agnostically learning a binary product distribution. Such a distribution is entirely determined by its coordinate-wise mean, which we denote by the vector p, and our first goal is to estimate p within ℓ_2-distance $\widetilde{O}(\varepsilon)$. We can borrow many of the ideas that we sketched in earlier applications of the filtering approach. Recall that the approach for robustly learning the mean of an identity covariance Gaussian was to compute the top absolute eigenvalue of a modified empirical covariance matrix. Our modification was crucially based on the promise that the covariance of the Gaussian is the identity. Here, it turns out that what we should do to modify the empirical covariance matrix is subtract off a diagonal matrix whose entries are p_i^2. These values seem challenging to directly estimate. Instead, we directly zero out the diagonal entries of the empirical covariance matrix. Then the filtering approach proceeds as before, and allows us to estimate p within ℓ_2-distance $\widetilde{O}(\varepsilon)$, as we wanted.

In the case when p has no coordinates that are too biased towards either zero or one, our estimate is already $\widetilde{O}(\varepsilon)$ close in total variation distance. However, when p has some very biased coordinates, this need not be the case. Each coordinate that is biased needs to be learned multiplicatively correctly. Nevertheless, we can use our estimate for p that is close in ℓ_2-distance as a starting point for handling binary product distributions that have imbalanced coordinates. Instead, we control the total variation distance via the χ^2-distance between the mean vectors. Let P and Q be two product distributions whose means are p and q respectively. Using the relationship between total variation distance and χ^2-distance, it follows that

$$d_{\mathrm{TV}}(P,Q)^2 \le 4 \sum_i \frac{(p_i - q_i)^2}{q_i(1-q_i)} \ .$$

So, if our estimate q is already close in ℓ_2-distance to p, we can interpret the right hand side above as giving a renormalization of how we should measure the distance between p and q so that being close (in χ^2-distance) implies that our estimate is close in total variation distance. We can then set up a corrected eigenvalue problem using our initial estimate q as follows. Let $\chi^2(v)_q = \sum_i v_i^2 q_i(1-q_i)$. Then, we compute

$$\max_{\chi^2(v)_q = 1} v^T \Sigma v \ ,$$

where Σ is the modified empirical covariance. In the full version of our paper, we show that this yields an estimate that is $\widetilde{O}(\sqrt{\varepsilon})$ close in total variation distance.

VII. Mixtures of Two Balanced Product Distributions

In this section, we study the problem of agnostically learning a mixture of two balanced binary product distributions. Let p and q be the coordinate-wise means of the two product distributions. Let $u = \frac{p}{2} - \frac{q}{2}$. Then, when there is no noise, the empirical covariance matrix is $\Sigma = uu^T + D$, where D is a diagonal matrix whose entries are $\frac{p_i+q_i}{2} - \frac{(p_i-q_i)^2}{4}$. Thus, it can already have a large eigenvalue. Now in the presence of corruptions it turns out that we can construct a filter when the *second* absolute eigenvalue is also large. But even if only the top absolute eigenvalue is large, we know that both p and q are close to the line $\mu + cv$, where μ is the empirical mean and v is the top eigenvector. And by performing a grid search over c, we will find a good candidate hypothesis.

Unfortunately, bounds on the top absolute eigenvalue do not translate as well into bounds on the total variation distance of our estimate to the true distribution, as they did in all previous cases (e.g., if the top absolute eigenvalue is small in the case of learning the mean of a Gaussian with identity covariance, we can just use the empirical mean, etc). In fact, an eigenvalue λ could just mean that p and q differ by $\sqrt{\lambda}$ along the direction v. However, we can proceed by zeroing out the diagonals. If uu^T has any large value along the diagonal, this operation can itself produce large eigenvalues. So, this strategy only works when $\|u\|_\infty$ is appropriately bounded. Moreover, there is a strategy to deal with large entries in u by guessing a coordinate whose value is large and conditioning on it, and once again setting up a modified eigenvalue problem. We defer the details to the full version of our paper. Our overall algorithm then follows from balancing all of these different cases.

VIII. Some Natural Approaches, and Why They Fail

In fact, the problem of agnostically learning a distribution in high-dimensions is so natural that in many of the settings, one would immediately wonder why simpler approaches do not work. Here we detail some other plausible approaches, and what causes them to lose dimension-dependent factors (if they have any guarantees at all!). For the discussion that follows, we note that by Fact III.1 in order to achieve an estimate that is $O(\varepsilon)$-close in total variation distance (for a Gaussian when μ is unknown and $\Sigma = I$) we require $\|\hat{\mu} - \mu\| = O(\varepsilon)$.

Learn Each Dimension Separately: Suppose we want to learn the mean of a Gaussian with covariance Σ. We could try to learn each coordinate of the mean separately, but since an ε-fraction of the samples are corrupted, our estimate can be off by ε in *each* coordinate and would be off by $\varepsilon\sqrt{d}$ in high dimensions.

Maximum Likelihood Estimator: The MLE is hard to compute, but even ignoring computational considerations it does not produce a robust estimate in the sense of Question I.1. It is well known [16], [17] that the MLE converges to the distribution $P' \in \mathcal{D}$ that is closest in KL-divergence to the distribution from which our samples were generated (i.e. after the adversary has added corruptions). However if an adversary places an ε-fraction of the points at some very large distance, then the estimate for the mean would need to move considerably in that direction. By placing the corruptions further and further away, the MLE can be an arbitrarily bad estimate.

Geometric Median: As we discussed, the Tukey depth [20] is one high-dimensional analogue of the median, but is hard to compute [21]. Another valid way to define the median in high dimensions is to set it to be the v that minimizes $\sum_{i=1}^{m} \|X_i - v\|_2$. In the full version of our paper, we show that this can also yield an estimate that is off by as much as $\varepsilon\sqrt{d}$ from the true mean.

Acknowledgment

ID supported in part by EPSRC grant EP/L021749/1, a Marie Curie Career Integration Grant, and a SICSA grant. GK supported by NSF Award CCF-0953960 (CAREER) and ONR grant N00014-12-1-0999. Part of this work was done while the author was an intern at Microsoft Research Cambridge and visiting the Simons Institute for the Theory of Computing. JL supported by NSF CAREER Award CCF-1453261, a Google Faculty Research Award, and an NSF Fellowship. Part of this work was done while the author was an intern at Microsoft Research Cambridge. JL would like to thank Michael B. Cohen and Samuel B. Hopkins for helpful discussions. AM supported by NSF CAREER Award CCF-1453261, a grant from the MIT NEC Corporation, and a Google Faculty Research Award. AS supported in part by EPSRC grant EP/L021749/1. Part of this work was performed while ID, DMK, and AS were at the University of Edinburgh.

References

[1] J. Feldman, R. O'Donnell, and R. Servedio, "Learning mixtures of product distributions over discrete domains," in *Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, 2005, pp. 501–510.

[2] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two Gaussians," in *STOC*, 2010, pp. 553–562.

[3] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *FOCS*, 2010, pp. 93–102.

[4] M. Belkin and K. Sinha, "Polynomial learning of distribution families," in *FOCS*, 2010, pp. 103–112.

[5] D. Hsu and S. M. Kakade, "Learning mixtures of spherical gaussians: moment methods and spectral decompositions," in *Innovations in Theoretical Computer Science, ITCS '13*, 2013, pp. 11–20.

[6] M. Cryan, L. Goldberg, and P. Goldberg, "Evolutionary trees can be learned in polynomial time in the two state general Markov model," *SIAM Journal on Computing*, vol. 31, no. 2, pp. 375–397, 2002.

[7] E. Mossel and S. Roch, "Learning nonsingular phylogenies and Hidden Markov Models," in *To appear in Proceedings of the 37th Annual Symposium on Theory of Computing (STOC)*, 2005.

[8] A. Anandkumar, D. Hsu, and S. Kakade, "A method of moments for mixture models and Hidden Markov Models," *Journal of Machine Learning Research - Proceedings Track*, vol. 23, pp. 33.1–33.34, 2012.

[9] S. Arora, R. Ge, and A. Moitra, "Learning topic models - going beyond SVD," in *FOCS 2012*, 2012, pp. 1–10.

[10] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade, "A tensor spectral approach to learning mixed membership community models," in *COLT 2013*, 2013, pp. 867–881.

[11] S. Arora, R. Ge, A. Moitra, and S. Sachdeva, "Provable ICA with unknown gaussian noise, and implications for gaussian mixtures and autoencoders," *Algorithmica*, vol. 72, no. 1, pp. 215–236, 2015.

[12] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, vol. 100, pp. 78–150, 1992.

[13] M. Kearns, R. Schapire, and L. Sellie, "Toward Efficient Agnostic Learning," *Machine Learning*, vol. 17, no. 2/3, pp. 115–141, 1994.

[14] L. Valiant, "Learning disjunctions of conjunctions," in *Proc. 9th IJCAI*, 1985, pp. 560–566.

[15] M. J. Kearns and M. Li, "Learning in the presence of malicious errors," *SIAM Journal on Computing*, vol. 22, no. 4, pp. 807–837, 1993.

[16] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, 1967, pp. 221–233. [Online]. Available: http://projecteuclid.org/euclid.bsmsp/1200512988

[17] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, no. 1, pp. 1–25, 1982.

[18] P. Huber and E. M. Ronchetti, *Robust statistics*. Wiley New York, 2009.

[19] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics. The approach based on influence functions*. Wiley New York, 1986.

[20] J. Tukey, "Mathematics and picturing of data," in *Proceedings of ICM*, vol. 6, 1975, pp. 523–531.

[21] D. S. Johnson and F. P. Preparata, “The densest hemisphere problem,” *Theoretical Computer Science*, vol. 6, pp. 93–107, 1978.

[22] T. Bernholt, “Robust estimators are hard to compute,” University of Dortmund, Germany, Tech. Rep., 2006.

[23] M. Hardt and A. Moitra, “Algorithms and hardness for robust subspace recovery,” in *COLT 2013*, 2013, pp. 354–375.

[24] P.-L. Loh and X. L. Tan, “High-dimensional robust precision matrix estimation: Cellwise corruption under -contamination,” 2015.

[25] S. Balmand and A. Dalalyan, “Convex programming approach to robust estimation of a multivariate gaussian model,” 2015.

[26] Y. Freund and Y. Mansour, “Estimating a mixture of two product distributions,” in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, ser. COLT ’99, 1999, pp. 53–62.

[27] R. Barlow, D. Bartholomew, J. Bremner, and H. Brunk, *Statistical Inference under Order Restrictions*. New York: Wiley, 1972.

[28] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 1985.

[29] B. W. Silverman, *Density Estimation*. London: Chapman and Hall, 1986.

[30] D. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley, 1992.

[31] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*. Springer: Springer Series in Statistics, 2001.

[32] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie, “On the learnability of discrete distributions,” in *Proc. 26th STOC*, 1994, pp. 273–282.

[33] Y. Freund and Y. Mansour, “Estimating a mixture of two product distributions,” in *Proceedings of the 12th Annual COLT*, 1999, pp. 183–192.

[34] S. Arora and R. Kannan, “Learning mixtures of arbitrary Gaussians,” in *Proceedings of the 33rd Symposium on Theory of Computing*, 2001, pp. 247–257.

[35] S. Vempala and G. Wang, “A spectral algorithm for learning mixtures of distributions,” in *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, 2002, pp. 113–122.

[36] J. Feldman, R. O’Donnell, and R. Servedio, “Learning mixtures of product distributions over discrete domains,” in *Proc. 46th IEEE FOCS*, 2005, pp. 501–510.

[37] C. Daskalakis, I. Diakonikolas, and R. Servedio, “Learning Poisson Binomial Distributions,” in *STOC*, 2012, pp. 709–728.

[38] C. Daskalakis, I. Diakonikolas, R. O’Donnell, R. Servedio, and L. Tan, “Learning Sums of Independent Integer Random Variables,” in *FOCS*, 2013, pp. 217–226.

[39] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun, “Efficient density estimation via piecewise polynomial approximation,” in *STOC*, 2014, pp. 604–613.

[40] ——, “Near-optimal density estimation in near-linear time using variable-width histograms,” in *NIPS*, 2014, pp. 1844–1852.

[41] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt, “Sample-optimal density estimation in nearly-linear time,” *CoRR*, vol. abs/1506.00671, 2015.

[42] C. Daskalakis, A. De, G. Kamath, and C. Tzamos, “A size-free CLT for poisson multinomials and its applications,” in *Proceedings of STOC’16*, 2016.

[43] I. Diakonikolas, D. M. Kane, and A. Stewart, “The fourier transform of poisson multinomial distributions and its algorithmic applications,” in *Proceedings of STOC’16*, 2016.

[44] R. Servedio, “Smooth boosting and learning with malicious noise,” in *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, 2001, pp. 473–489.

[45] ——, “Smooth boosting and learning with malicious noise,” *JMLR*, vol. 4, pp. 633–648, 2003.

[46] A. Klivans, P. Long, and R. Servedio, “Learning halfspaces with malicious noise,” 2009, to appear in *Proc. 17th Internat. Colloq. on Algorithms, Languages and Programming (ICALP)*.

[47] P. Awasthi, M. F. Balcan, and P. M. Long, “The power of localization for efficiently learning linear separators with noise,” in *STOC 2014*, 2014, pp. 449–458.

[48] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *J. ACM*, vol. 58, no. 3, p. 11, 2011.

[49] T. Zhang and G. Lerman, “A novel M-estimator for robust PCA,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 749–808, Jan. 2014.

[50] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, “Robust computation of linear models, or how to find a needle in a haystack,” *CoRR*, vol. abs/1202.4044, 2012. [Online]. Available: http://arxiv.org/abs/1202.4044

[51] S. C. Brubaker, “Robust PCA and clustering in noisy mixtures,” in *SODA 2009*, 2009, pp. 1078–1087.

[52] K. A. Lai, A. B. Rao, and S. Vempala, “Agnostic estimation of mean and covariance,” in *Proceedings of FOCS’16*, 2016.

[53] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour, “Near-optimal-sample estimators for spherical gaussian mixtures,” in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 1395–1403.