Agnostic Estimation of Mean and Covariance

Kevin A. Lai, Anup B. Rao and Santosh Vempala
School of Computer Science
Georgia Tech.
Atlanta, U.S.A.
$\{$*kevinlai, anup.rao, vempala*$\}$*@gatech.edu*

***Abstract*—We consider the problem of estimating the mean and covariance of a distribution from i.i.d. samples in the presence of a fraction of *malicious* noise. This is in contrast to much recent work where the noise itself is assumed to be from a distribution of known type. The agnostic problem includes many interesting special cases, e.g., learning the parameters of a single Gaussian (or finding the best-fit Gaussian) when a fraction of data is adversarially corrupted, agnostically learning mixtures, agnostic ICA, etc. We present polynomial-time algorithms to estimate the mean and covariance with error guarantees in terms of information-theoretic lower bounds. As a corollary, we also obtain an agnostic algorithm for Singular Value Decomposition.**

***Keywords*-Mean estimation; covariance; PCA; agnostic learning; robust statistics.**

I. Introduction

The mean and covariance of a probability distribution are its most basic parameters (if they are bounded). Many families of distributions are defined using only these parameters. Estimating the mean and covariance from iid samples is thus a fundamental and classical problem in statistics. The sample mean and sample covariance are generally the best possible estimators (under mild conditions on the distribution such as their existence). However, they are highly sensitive to noise. The main goal of this paper is to estimate the mean, covariance and related functions in spite of arbitrary (adversarial) noise.

Methods for efficient estimation, in terms of sample complexity and time complexity, play an important role in many algorithms. One such class of problems is unsupervised learning of generative models. Here the input data is assumed to be iid from an unknown distribution of a known type. The classical instantiation is Gaussian mixture models, but many other models have been studied widely. These include topic models, stochastic block models, Independent Component Analysis (ICA) etc. In all these cases, the problem is to estimate the parameters of the underlying distribution from samples. For example, for a mixture of k Gaussians in $\mathbb{R}^n$, it is known that the sample and time complexity are bounded by $n^{O(k)}$ in general [1], [2], [3] and by poly(n, k) under natural separation assumptions [4], [5], [6], [7], [8], [9], [10]. For ICA, samples are of the form Ax where A is unknown and x is chosen randomly from an unknown (non-Gaussian) product distribution; the problem is to estimate the linear transformation A and thus unravel the underlying product structure [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. These, and other models (see e.g., [21]), have been a rich and active subject of study in recent years and have lead to interesting algorithms and analyses.

The Achilles heel of algorithms for generative models is the assumption that data is *exactly* from the model. This is crucial for known guarantees, and relaxations of it are few and specialized, e.g., in ICA, data could by noisy, but the noise itself is assumed to be Gaussian. Assumptions about rank and sparsity are made in a technique that is now called Robust PCA [22], [23], [24]. There have been attempts [25], [26] at achieving robustness by L1 minimization, but they don't give any error bounds on the output produced. A natural, important and wide open problem is estimating the parameters of generative models in the presence of arbitrary, i.e., *malicious* noise, a setting usually referred to as *agnostic* learning. The simplest version of this problem is to *estimate a single Gaussian* in the presence of malicious noise. Alternatively, this can be posed as the problem of finding a best-fit Gaussian to data or agnostically learning a single Gaussian. We consider the following generalization:

Problem 1 [Mean and Covariance]: Given points in $\mathbb{R}^n$ *that are each, with probability* $1 - \eta$ *from an unknown distribution with mean* μ *and covariance* Σ*, and with probability* η *completely arbitrary, estimate* μ *and* Σ*.*

There is a large literature on *robust* statistics (see e.g., [27], [28], [29]), with the goal of finding estimators that are stable under perturbations of the data. The classic example for points on a line is that the sample median is a robust estimator while the sample mean is not (a single data point can change the mean arbitrarily). One measure for robustness of an estimator is called *breakdown* point, which is the minimum fraction of noise that can make the estimator arbitrarily bad. Robust statistics have been proposed and studied for mean and covariance estimation in high dimension as well (see [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40] and the references therein). Most commonly used methods (including M-estimators) to estimate the covariance matrix were shown to have very low break down points [34]. The notion of robustness we

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.76

consider quantifies how far the estimated value is from the true value. To the best of our knowledge, all the papers either suffer from the difficulty that their algorithms are computationally very expensive, namely exponential time in the dimension, or have poor or no guarantees for the output. Tukey's median [31]) is an example of the former. It is defined as the *deepest* point with respect to a given set of points $\{\boldsymbol{x}_i\}_i$. As proven in [40], this is an optimal estimate of the mean. But there is no known polynomial time algorithm to compute this. Another well-known proposal (see [41]) is the geometric median:

$$\arg\min_{\boldsymbol{y}} \sum_i \|\boldsymbol{y}-\boldsymbol{x}_i\|_2.$$

This has the advantage that it can be computed via a convex program. Unfortunately, as we observe here (see Proposition II.1), the error of the mean estimate produced by this method grows polynomially with the dimension (also see [42]).

This leads to the question, what is the best approximation one can hope for with η arbitrary (adversarial) noise. From a purely information-theoretic point of view, it is not hard to see that even for a single Gaussian $N(\mu, \sigma^2)$ in one dimension, the best possible estimation of the mean will have error as large as $\Omega(\eta\sigma)$, i.e., any estimate $\tilde{\mu}$ can be forced to have $\|\mu-\tilde{\mu}\| = \Omega(\eta\sigma)$. For a more general distribution, this can be slightly worse, namely, $\Omega(\eta^{3/4}\sigma)$ (see Section II-A). What about in $\mathbb{R}^n$? Perhaps surprisingly, but without much difficulty, one can show that the information-theoretic upper bound matches the lower bound in any dimension, with no dependence on the dimension. This raises a compelling algorithmic question: what are the best estimates for the mean and covariance that can be computed efficiently?

In this paper, we give polynomial time algorithms to estimate the mean with error that is close to the information-theoretically optimal estimator. The dependence on the dimension, of the error in the estimated mean, is only $\sqrt{\log n}$. To the best of our knowledge, this is the first polynomial-time algorithm with an error dependence on dimension that is less than $\sqrt{n}$, the bound achieved by the geometric median. Moreover, as we state precisely later, our techniques extend to very general input distributions and to estimating higher moments.

Our algorithm is practical. A matlab implementation for mean estimation can be found in [43]. It takes less a couple of seconds to run on a 500-dimensional problem with 5000 samples on a personal laptop.

A. Model

We are given points $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \in \mathbb{R}^n$ sampled according to the following rule. With $1-\eta$ probability each $\boldsymbol{x}_i$ is independently sampled from a distribution $\mathcal{D}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and with η probability it is picked by an adversary. For ease of notation, we will write $\boldsymbol{x}_i \sim \mathcal{D}_\eta$ when we want to say the $\boldsymbol{x}_i$ is picked according to the above rule. The problem we are interested in is to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given the samples. In the following, we will consider mainly two kinds of distributions.

Gaussian: $\mathcal{D} = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Bounded Moments: Let $\mathcal{D}$ is a distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We say it has bounded $2k$'th moments if there exists a constant C_{2k} such that for every unit vector $\boldsymbol{v}$,

$$\begin{aligned} \mathbf{E}\left((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{v}\right)^{2k} &\le C_{2k}\left(\mathbf{E}\left((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{v}\right)^2\right)^k \qquad (1)\\ &= C_{2k}(\mathbf{Var}\left[\boldsymbol{x}^T\boldsymbol{v}\right])^k. \qquad (2) \end{aligned}$$

Here $\mathbf{Var}\left[\boldsymbol{x}^T\boldsymbol{v}\right] = \left(\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}\right)^2$ is the variance of $\boldsymbol{x}$ along $\boldsymbol{v}$. For mean estimation, C_4 will be used, and for covariance estimation, C_8 will be needed.

B. Main Results

All the results we state hold with probability $1 - 1/\operatorname{poly}(n)$ unless otherwise mentioned. We will also assume η is a less than a universal constant. We begin with agnostic mean estimation.

Theorem I.1 (Gaussian mean)**.** *Let $\mathcal{D} = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^n$. There exists a $\operatorname{poly}(n, 1/\epsilon)$-time algorithm that takes as input $m = O\left(\frac{n(\log n + \log 1/\epsilon)\log n}{\epsilon^2}\right)$ independent samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \sim \mathcal{D}_\eta$ and computes $\widehat{\boldsymbol{\mu}}$ such that the error $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_2$ is bounded as follows:*

$$\begin{array}{ll} O\left(\eta+\epsilon\right)\sigma\sqrt{\log n} & \textit{if } \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I} \\ O\left(\eta^{1/2}+\epsilon\right)\|\boldsymbol{\Sigma}\|_2^{1/2}\log^{1/2} n & \textit{otherwise.} \end{array}$$

We note that the sample complexity is nearly linear, and almost matches the complexity for mean estimation with no noise.

Remark I.2. If we take $m = O\left(\frac{n^2(\log n + \log 1/\eta)\log n}{\eta^2}\right)$ samples, and assume that $\eta < c/\log n$ for a small enough constant $c > 0$, then by combining theorems I.5 and I.1, we can improve the η dependence for the non-spherical Gaussian case in Theorem I.1 to $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_2 = O\left(\eta^{3/4}\right)\|\boldsymbol{\Sigma}\|_2^{1/2}\log^{1/2} n$.

Our next theorem is a similar result for much more general distributions.

Theorem I.3 (General mean)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^n$ with mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$, and bounded fourth moments (see Equation 1). There exists a $\operatorname{poly}(n, 1/\epsilon)$-time algorithm that takes as input a parameter η and $m = O\left(\frac{n(\log n + \log 1/\epsilon)\log n}{\epsilon^2}\right)$ independent samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \sim \mathcal{D}_\eta$, and computes $\widehat{\boldsymbol{\mu}}$ such that the error $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_2$ is bounded* *as follows:*

$$\begin{aligned}&O\left(C_4^{1/4}(\eta+\epsilon)^{3/4}\right)\sigma\sqrt{\log n} && \textit{if } \boldsymbol{\Sigma}=\sigma^2\boldsymbol{I}\\ &O\left(\eta^{1/2}+C_4^{1/4}(\eta+\epsilon)^{3/4}\right)\|\boldsymbol{\Sigma}\|_2^{1/2}\log^{1/2} n && \textit{otherwise.}\end{aligned}$$

The bounds above are nearly the best possible (up to a factor of $O(\sqrt{\log n})$) when the covariance is a multiple of the identity.

Observation I.4 (Lower Bounds)**.** *Let $\mathcal{D}$ be a distribution with mean $\boldsymbol{\mu}\in\mathbb{R}^n$ and covariance $\boldsymbol{\Sigma}$. Any algorithm that takes m (not necessarily $O(\mathrm{poly}(n))$) samples $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\sim\mathcal{D}_\eta$, and computes a $\widehat{\boldsymbol{\mu}}$ should have with constant probability the error $\|\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}\|_2$ is*

$$\begin{aligned}&\Omega(\eta\sqrt{\|\boldsymbol{\Sigma}\|_2}) && \textit{if } \mathcal{D}=N(\boldsymbol{\mu},\boldsymbol{\Sigma})\\ &\Omega(\eta^{3/4}\sqrt{\|\boldsymbol{\Sigma}\|_2}) && \textit{if } \mathcal{D} \textit{ has bounded fourth moments.}\end{aligned}$$

Theorem I.5 (Covariance Estimation)**.** *Let $\mathcal{D}$ be a distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and that (a) for $\boldsymbol{x}\sim\mathcal{D}$, $\boldsymbol{x}$ and $(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T$ have bounded fourth moments with constants C_4 and $C_{4,2}$(see Equation 1) respectively. (b) $\mathcal{D}$ is an (unknown) affine transformation of a 4-wise independent distribution. Then, there is an algorithm that takes as input $m=O\left(\frac{n^2(\log n+\log 1/\epsilon)\log n}{\epsilon^2}\right)$ samples $\boldsymbol{x}_1,\ldots\boldsymbol{x}_m\sim\mathcal{D}_\eta$ and η and computes in $\mathrm{poly}(n,1/\epsilon)$-time a covariance estimate $\widehat{\boldsymbol{\Sigma}}$ such that*

$$\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_F=O\left(\eta^{1/2}+C_{4,2}^{1/4}(\eta+\epsilon)^{3/4}\right)C_4^{1/2}\|\boldsymbol{\Sigma}\|_2\log^{1/2} n$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

If $\mathcal{D}=N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then it satisfies the hypothesis of the above theorem. More generally, it holds for any 8-wise independent distribution with bounded eighth moments and whose fourth moment along any direction is at least $(1+c)$ times the square of the second moment for some $c>0$. We also note that if the distribution is isotropic, then covariance estimation is essentially a 1-d problem and we get a better bound.

Theorem I.6 (Agnostic 2-norm)**.** *Suppose $\mathcal{D}$ is a distribution which satisfies the following concentration inequality: there exists a constant γ such that for every unit vector $\boldsymbol{v}$*

$$\Pr\left(\left|(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{v}\right|>t\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}\right)\le e^{-t^\gamma}.$$

Then, there is an algorithm that runs in $\mathrm{poly}(n,1/\eta)$ time that takes as input η and $m=O\left(\frac{n^3(\log n/\eta)^2\log n}{\eta^2}\right)$ independent samples $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\sim\mathcal{D}_\eta$, and computes $\widehat{\lambda}_{\max}$ such that

$$(1-O(\eta))\,\|\boldsymbol{\Sigma}\|_2\le\widehat{\lambda}_{\max}\le\left(1+O(\eta\log^{2/\gamma}n/\eta)\right)\|\boldsymbol{\Sigma}\|_2.$$

In independent work, [44] gave a similar algorithm, which they call a Gaussian filtering method, for agnostic mean estimation assuming a spherical covariance matrix; while their guarantees are specifically for Gaussians, the error term in their guarantee grows only with $\log(1/\eta)$ rather than $\log n$. They also give a completely different algorithm based on the Ellipsoid method, for a simple family of distributions including Gaussian and Bernoulli.

As a corollary of Theorem I.5, we get a guarantee for agnostic SVD.

Theorem I.7 (Agnostic SVD)**.** *Let $\mathcal{D}$ is a distribution that satisfies the hypothesis of Theorem I.5. Let $\boldsymbol{\Sigma}_k$ be the best rank k approximation to $\boldsymbol{\Sigma}$ in $\|\cdot\|_F$ norm. There exists a polynomial time algorithm that takes as input η and $m=\mathrm{poly}(n)$ samples from $\mathcal{D}_\eta$. It produces a rank k matrix $\widehat{\boldsymbol{\Sigma}}_k$ such that*

$$\left\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_k\right\|_F\le\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_k\|_F+O\left(\sqrt{\eta\log n}\right)\|\boldsymbol{\Sigma}\|_2.$$

Given the wide applicability of SVD to data, we expect the above theorem will have many applications. As an illustration, we derive a guarantee for agnostic Independent Component Analysis (ICA). In standard ICA, input data points x are generated as As with a fixed unknown $n\times n$ full-rank matrix A and s generated from an unknown product distribution with non-Gaussian components. The problem is to estimate the matrix A (the "basis") from a polynomial number of samples in polytime. There is a large literature of algorithms for this problem and its extensions [11], [12], [13], [14], [15], [16], [17], [18], [19]. However, all these algorithms rely on no noise or the noise being random (typically Gaussian) and require estimating singular values to within $1/\mathrm{poly}(n)$ accuracy, and therefore unable to handle adversarial noise. On the other hand, the algorithm from [20], which gives a sample complexity of $\tilde{O}(n)$, only requires estimating singular values to within $1/\mathrm{poly}(\log n)$. Our algorithm for agnostic SVD together with the Recursive Fourier PCA algorithm of [20] results in an efficient algorithm for agnostic ICA, tolerating noise $\eta=O(1/\log^c n)$ for a fixed constant c. To the best of our knowledge, this is the first polynomial-time algorithm that can handle more than an inverse $\mathrm{poly}(n)$ amount of noise.

Theorem I.8 (Agnostic Standard ICA)**.** *Let $x\in\mathbb{R}^n$ be given by a noisy ICA model $x=As$ with probability $1-\eta$ and be arbitrary with probability η, where $A\in\mathbb{R}^{n\times n}$ has condition number κ, the components of s are independent, $\|s\|\le K\sqrt{n}$ almost surely, and for each i, $\mathbf{E}s_i=0,\mathbf{E}s_i^2=1$, $|\mathbf{E}|s_i|^4-3|\ge\Delta$ and $\max_i\mathbf{E}|s_i|^5\le M$. Then for any $\epsilon<\Delta^3/(10^8M^2\log^3 n),1/(\kappa^4\log n)$ and $\eta<\epsilon/2$, there is an algorithm that, with high probability, finds vectors $\{b_1,\ldots,b_n\}$ such that there exist signs $\xi_i=\pm1$ satisfying $\left\|A^{(i)}-\xi_ib_i\right\|\le\epsilon\|A\|_2$ for each column $A^{(i)}$ of A, using $\mathrm{poly}(n,K,\Delta,M,\kappa,\frac{1}{\epsilon})$ samples. The running time is bounded by the time to compute $\tilde{O}(n)$ SVDs on real symmetric matrices of size $n\times n$.*

Our results can also be used to estimate the mean and covariance of noisy Bernoulli product distributions, i.e.

distributions in which each coordinate i is 1 with probability p_i and 0 with probability $1 - p_i$. In one dimension, C_4 for a Bernoulli distribution is $\frac{(1-p)^2}{p} + \frac{p^2}{1-p}$. For a Bernoulli product distribution, C_4 will be within a constant of $\max_i \left\{ \frac{(1-p_i)^2}{p_i} + \frac{p_i^2}{1-p_i} \right\}$. Then Theorem I.3 can be applied to get an estimate $\widehat{\boldsymbol{\mu}}$ for the mean. For instance, if $\forall i, p_i = p$ and $p \geq \frac{1}{2}$, then $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_2 = O\left(\sqrt{\eta(1 + \sqrt{\eta p})p \log n}\right)$. If C_4 is constant, then by Theorem I.5, we can get an estimate for the covariance. Detailed proofs of all the theorems can be found in the full version of the paper available at http://arxiv.org/abs/1604.06968.

II. Main Ideas and Algorithms

Here we discuss the key ideas of the algorithms. The algorithm AgnosticMean (Algorithm 3) alternates between an outlier removal step and projection onto the top $n/2$ principal components; these steps are repeated. It is inspired by the work of Brubaker [45] who gave an agnostic algorithm for learning a mixture of well-separated spherical Gaussians.

For illustration, let us assume for now that the underlying distribution is $\mathcal{D} = N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$. We are given a set S of $m = \text{poly}(n)$ points from $\mathcal{D}_\eta$, and $S = S_G \cup S_N$ be the points sampled from the Gaussian and the adversary respectively. Let us also assume that $|S_N| = \eta|S|$. We will use the notation $\boldsymbol{\mu}_T$ for mean of the points in a set T, and $\boldsymbol{\Sigma}_T$ for covariance of the points in T. We then have

$$\boldsymbol{\Sigma}_S = (1-\eta)\sigma^2 \boldsymbol{I} + \eta \boldsymbol{\Sigma}_{S_N} + \eta(1-\eta)(\boldsymbol{\mu}_S - \boldsymbol{\mu}_N)(\boldsymbol{\mu}_S - \boldsymbol{\mu}_N)^T. \tag{3}$$

If the dimension is $n = 1$, then we can show that the median of S is an estimate for $\boldsymbol{\mu}$ correct up to an additive error of $O(\eta\sigma)$. Even if we just knew the direction of the *mean shift* $\boldsymbol{\mu}_S - \boldsymbol{\mu} = \eta(\boldsymbol{\mu}_G - \boldsymbol{\mu}_N)$, then we can estimate $\boldsymbol{\mu}$ by first projecting the sample S on the line along $\boldsymbol{\mu} - \boldsymbol{\mu}_S$ and then finding the median. This would give an estimator $\widehat{\boldsymbol{\mu}}$ satisfying $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 = O(\eta\sigma)$. So we can focus on finding the direction of $\boldsymbol{\mu}_S - \boldsymbol{\mu}$. One would guess that the top principal component of the covariance matrix of S would be a good candidate. But it is easy for the adversary to choose S_N to make this completely useless. Since the noise points S_N can be anything, just two points from S_N placed far away on either side of the mean $\boldsymbol{\mu}$ along a particular line passing through $\boldsymbol{\mu}$ are sufficient to make the variance in that direction blow up arbitrarily. But we can limit this effect to some extent by an outlier removal step. By a standard concentration inequality for Gaussians, we know that the points in S_G lie in a ball of radius $O(\sigma\sqrt{n})$ around the mean. So, if we can just find a point inside or close to the convex hull of the Gaussian and throw away all the points that lie outside a ball of radius $C\sigma\sqrt{n}$ around this point, we preserve all the points in S_G. This will also contain the effect of noise points on the variance since now they are restricted to be within $O(\sigma\sqrt{n})$ distance of $\boldsymbol{\mu}$. We will see later that we can use coordinate-wise median as the center of the ball. By computing the variance by projecting onto any direction, we can figure out σ^2 up to a $1 \pm O(\eta)$ factor. From now on, we assume that all points in S lie within a ball of radius $O(\sigma\sqrt{n})$ centered at $\boldsymbol{\mu}$.

But even after this restriction, the top principal component may not contain any information about the mean shift direction. By just placing (say) $\eta/10$ noise points along the e_1 direction at $\pm\sigma\sqrt{n}$, and all the remaining noise points perpendicular to this at a single point at a smaller distance, we can make e_1 the top principal component. But e_1 is perpendicular to the mean shift direction.

The idea to get around this is that even if the top principal component of $\boldsymbol{\Sigma}_S$ may not be along the mean-shift direction, the span (call it V) of top $n/2$ principal components of $\boldsymbol{\Sigma}_S$ will contain a big projection of the mean-shift vector. This is because, if a big component of the the mean-shift vector was in the span (say W) of bottom $n/2$ principal components of $\boldsymbol{\Sigma}_S$, by Equation 3 this would mean that there is a vector in W with a large Rayleigh quotient. This implies that the top $n/2$ eigenvalues of $\boldsymbol{\Sigma}_S$ are all big. Since $\boldsymbol{\Sigma}_S = (1-\eta)\sigma^2\boldsymbol{I} + \boldsymbol{A}$, where $\boldsymbol{A} = \eta\boldsymbol{\Sigma}_{S_N} + \eta(1-\eta)(\boldsymbol{\mu}_S - \boldsymbol{\mu}_N)(\boldsymbol{\mu}_S - \boldsymbol{\mu}_N)^T$, this is possible only if $\text{Tr}(\boldsymbol{A})$ is large. But since the distance of each point in S from $\boldsymbol{\mu}$ is $O(\sigma\sqrt{n})$, the trace of $\boldsymbol{A}$ cannot be too large. Therefore, in the space W, we can just compute the sample mean $\boldsymbol{P}_W\boldsymbol{\mu}_S$ and it will be close to $\boldsymbol{P}_W\boldsymbol{\mu}$. We still have to find the mean in the space V. But we do this by recursing the above procedure in V. At the end we will be left with a one-dimensional space, and then we can just find the median. This recursive projection onto the top $n/2$ principal components is done in Algorithm 3 .

This generalizes to the non-spherical Gaussians with a few modifications. We use a different outlier removal step. In the non-spherical case, it is not trivial to compute $\|\boldsymbol{\Sigma}\|_2$ to be used as the radius of the ball. We give an algorithm for this later on. To limit the effect of noise, we use a damping function. Instead of discarding points outside a certain radius, we damp every point by a weight so that further away points get lower weights. This is done in OutlierDamping (Algorithm 1). We get the guarantees of Theorem I.1 by running AgnosticMean (Algorithm 3) with the outlier removal routine being OutlierDamping.

We then turn to more general distributions which have bounded fourth moments. We need bounded fourth moments to ensure that the mean and covariance matrix of the distribution $\mathcal{D}$ do not change much even after conditioning by an event that occurs with probability $1 - \eta$. One difficulty for general distributions is that the outlier damping doesn't work. So for distributions $\mathcal{D}$ with bounded fourth moments, we have another outlier removal routine called OutlierTruncation$(\cdot, \eta)$. In this routine, we first find a point analogous to the coordinate-wise median for the Gaussians, and then consider a ball big enough to contain $1 - \eta$ fraction of S. We throw away all the points outside this ball. We get the guarantees of Theorem I.3 by running AGNOSTICMEAN (Algorithm 3) with the outlier removal routine being OUTLIERTRUNCATION (Algorithm 2).

We now have an algorithm to estimate the mean of very general (with bounded fourth moments) distributions. To estimate the covariance matrix, we observed that the covariance matrix of a distribution $\mathcal{D}$ is given by $\mathbf{E}_{\mathcal{D}}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T$. If we knew what $\boldsymbol{\mu}$ was, then covariance can be computed by estimating the mean of the second moments. To compute the mean of the second moments, we can treat $(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T$ as a vector in n^2 dimensions and run the algorithm for mean estimation. Also, we can estimate $\boldsymbol{\mu}$ by the same algorithm. Therefore, we get Theorem I.5 by running COVARIANCEESTIMATION (Algorithm 4).

Algorithm AGNOSTICOPERATORNORM (Algorithm 5) estimates the 2-norm $\|\boldsymbol{\Sigma}\|_2$ for general distributions. For illustration, suppose $\mathcal{D} = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and we are given $m = \text{poly}(n)$ samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \sim \mathcal{D}_\eta$, and the mean $\boldsymbol{\mu}$. We consider the covariance-like matrix

$$\boldsymbol{\Sigma}(S, \boldsymbol{\mu}) = \frac{1}{m}\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T.$$

Since $1-\eta$ fraction of the points in S are from the Gaussian, we have $\boldsymbol{\Sigma}(S,\boldsymbol{\mu}) \succeq (1-\eta)\boldsymbol{\Sigma}$. Therefore, the top eigenvalue σ^2 of $\boldsymbol{\Sigma}(S,\boldsymbol{\mu})$ is at least $(1-\eta)\|\boldsymbol{\Sigma}\|_2$. Let $\boldsymbol{v}$ be the top eigenvector of $\boldsymbol{\Sigma}(S,\boldsymbol{\mu})$. If the Gaussian variance along $\boldsymbol{v}$ (which can be computed up to $1\pm\eta$ factor) is much less than σ^2, this should be because there are a lot of noise points in S whose projections onto $\boldsymbol{v}$ are big compared to the projection of Gaussian points in S. We remove points in S that have big projection and then iterate the entire procedure. We later show that this procedure terminates in $\text{poly}(n)$ steps and when it terminates the top eigenvalue of $\boldsymbol{\Sigma}(S,\boldsymbol{\mu})$ is close to that of $\boldsymbol{\Sigma}$.

Theorem I.7 follows easily from Theorem I.5. Let $\widehat{\boldsymbol{\Sigma}}_k$ be the top-k eigenspace of $\widehat{\boldsymbol{\Sigma}}$ from Theorem I.5. We then have

$$\begin{aligned}\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_k\right\|_F &\overset{(a)}{\le} \left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\|_F + \left\|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_k\right\|_F \\ &\overset{(b)}{\le} \left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\|_F + \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_k\right\|_F \\ &\overset{(c)}{\le} 2\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\|_F + \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_k\|_F \\ &\overset{(d)}{\le} \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_k\|_F + O\left(\sqrt{\eta \log n}\right)\|\boldsymbol{\Sigma}\|_2.\end{aligned}$$

$(a), (c)$ follow from triangle inequality, (b) follows from the fact that $\widehat{\boldsymbol{\Sigma}}_k$ is the best rank-k approximation and (d) from the guarantees of Theorem I.5.

Finally we outline the application to agnostic ICA. The algorithm from [20]. Proceeds by first estimating the mean and covariance, in order to make the underlying distribution isotropic. Here we estimate the covariance matrix $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}$ and use it to determine a new isotropic transformation $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$. Since our agnostic SVD algorithm gives a guarantee of $\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F \le O(\sqrt{\nu \log n})\|\boldsymbol{\Sigma}\|_2$, the isotropic transformation results in a guarantee of

$$\|\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - I\|_2 \le O(\sqrt{\eta \log n})\frac{\|\boldsymbol{\Sigma}\|_2}{\|\boldsymbol{\Sigma}^{-1}\|_2} = O(\sqrt{\eta \log n}\kappa^2).$$

Next the algorithm estimates a weighted covariance matrix $\boldsymbol{W}$ with the weight of a point $\boldsymbol{x}$ proportional to $\cos(\boldsymbol{u}^T\boldsymbol{x})$ for $\boldsymbol{u}$ chosen from a Gaussian distribution; it computes the SVD of $\boldsymbol{W}$. For this we use our algorithm again (the weights are applied individually to each sample). The main guarantee is that the eigenvectors of this weighted covariance approximate the columns of A. This relies on the maximum eigenvalue gap of $\boldsymbol{W}$ being large, and it has to be approximated to within additive error $\epsilon = O(1/(\log n)^3)$. Theorem I.7 implies that the additional error in eigenvalues is bounded by $O(\sqrt{\eta \log n})\|\boldsymbol{\Sigma}\|_2$, and therefore it suffices to have $\sqrt{\eta \log n} < c/(\log n)^3$ for a sufficiently small constant c that depends only on the cumulant and moment bound assumptions (i.e., Δ, M). Thus, if suffices to have $\eta < \epsilon/2 \le c(\log n)^{-7}$.

A. Lower Bounds: Observation I.4

In this section we will show the lower bounds stated in Observation I.4. For Gaussian distributions, this is a special case of a theorem proved in [40]. We reproduce the relevant part here for completeness. We will show that there are distributions $\mathcal{D}_1 = N(\boldsymbol{\mu}_1, \sigma^2\boldsymbol{I}), \mathcal{D}_2 = N(\boldsymbol{\mu}_2, \sigma^2\boldsymbol{I})$ and distributions Q_1, Q_2 such that $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 = \Omega(\eta\sigma)$ and

$$\mathcal{D}_\eta = (1-\eta)\mathcal{D}_1 + \eta Q_1 = (1-\eta)\mathcal{D}_2 + \eta Q_2. \qquad (4)$$

So, given $\mathcal{D}_\eta$, no algorithm can distinguish between $\mathcal{D}_1, \mathcal{D}_2$. Let ϕ_1 be p.d.f of $\mathcal{D}_1$ and ϕ_2 be the p.d.f of $\mathcal{D}_2$. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ be such that the total variation distance between $\mathcal{D}_1, \mathcal{D}_2$ is

$$\frac{1}{2}\int |\phi_1 - \phi_2| dx = \frac{\eta}{1-\eta}.$$

By a standard inequality for the total variation distance of Gaussian distributions, this implies that $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 \ge \frac{2\eta\sigma}{1-\eta}$. Let Q_1 be the distribution with p.d.f $\frac{1-\eta}{\eta}(\phi_2 - \phi_1)\mathbf{1}_{\phi_2 \ge \phi_1}$ and Q_2 be the distribution with p.d.f $\frac{1-\eta}{\eta}(\phi_1 - \phi_2)\mathbf{1}_{\phi_1 \ge \phi_2}$. It is now easy to verify that Equation 4 is satisfied. This proves item one of Observation I.4.

For the distributions with bounded fourth moments, consider the following two one-dimensional distributions. $\mathcal{D}_1$ is supported on two points $\{-\sigma, \sigma\}$ with the corresponding probabilities $\{1/2, 1/2\}$. $\mathcal{D}_2$ is supported on three points $\{-\sigma, \sigma, \sigma/\eta^{1/4}\}$ with probabilities $\{(1-\eta)/2, (1-\eta)/2, \eta\}$ respectively. Let $\eta \le 1/4$. It is easy to check that both $\mathcal{D}_1$ and $\mathcal{D}_2$ have bounded fourth moments with the constant $C_4 = 8$. Furthermore, $\mathcal{D}_2$ can be obtained from $\mathcal{D}_1$ by adding η fraction of noise points. So no algorithm can distinguish between the two distributions. Since their means differ by $\eta^{3/4}\sigma$, no algorithm can get an estimate better than this.

We will now show that the geometric median:

$$\arg\min_{\boldsymbol{y}} \sum_i \|\boldsymbol{x}_i - \boldsymbol{y}\|_2$$

has a $\sqrt{n}$ dependence on the dimension. We show this in the Gaussian case even if we have access to the whole distribution, but with η fraction of noise points placed all at a single point far away from most of the Gaussian points.

Proposition II.1 (Geometric Median). *Let $\mathcal{D} = N(\mathbf{0}, \boldsymbol{\Sigma})$ be a distribution with diagonal covariance matrix $\boldsymbol{\Sigma}$ whose variance along the coordinate direction $\boldsymbol{e}_1$ is zero, and equal to 1 in all the other coordinate directions. Assume there is an η fraction of noise at a distance $a = n$ along $\boldsymbol{e}_1$. Let*

$$t_0 = \arg\min_t (1-\eta)\mathbf{E}_{\boldsymbol{x}\sim\mathcal{D}}\left(\sqrt{t^2 + x_2^2 + ... + x_n^2}\right) + \eta(a-t). \tag{5}$$

Then, $t = \Omega(\eta\sqrt{n})$.

Proof: We have that at the minimizer t_0, the derivative with respect to t is zero. Therefore, we should have

$$\mathbf{E}_{\boldsymbol{x}\sim\mathcal{D}} \frac{t_0}{\sqrt{t_0^2 + x_2^2 + ... + x_n^2}} = \frac{\eta}{1-\eta}.$$

Consider $f(t) = \mathbf{E}_{\boldsymbol{x}\sim\mathcal{D}} \frac{t}{\sqrt{t^2+x_2^2+...+x_n^2}}$. It is clear from Equation 5 that $t_0 > 0$. We claim that if $t = \alpha\eta\sqrt{n}$ for a small enough constant α, then $f(t) \le \frac{\eta}{1-\eta}$. Suppose $t_1 = \alpha\eta\sqrt{n}$. Since $\boldsymbol{x} \sim \mathcal{D}$, $\|\boldsymbol{x}\|_2^2 \ge n/2$ with exponential probability. Therefore,

$$\begin{aligned} f(t_1) &\le \mathbf{E}_{\boldsymbol{x}\sim\mathcal{D}} \frac{t_1}{\sqrt{t_1^2 + n/2}} \\ &\le \frac{t_1\sqrt{2\pi}}{\sqrt{t_1^2 + n/2}} \le \alpha\eta\sqrt{2\pi}. \end{aligned}$$

The claim, and hence the proof follows.

□

B. Algorithms

Our algorithms are based on outlier removal and SVD. To simplify the proofs, we use new samples for each step of the algorithm. The total sample complexity is given in the theorems.

1) Outlier Removal: For outlier removal, we use one of the following two simple routines. The first, which we call *OutlierDamping*, returns a vector of positive weights, one for each sample point.

Algorithm 1: OUTLIERDAMPING(S)

```
Input: $S \subset \mathbb{R}^n$ with $|S| = m$
Output: $S \subset \mathbb{R}^n, \boldsymbol{w} = (w_1, ..., w_m) \in \mathbb{R}^m$
  1) if $n = 1$:
     Return $(S, -1)$.
  2) Let $\boldsymbol{a}$ be the coordinate-wise median of $S$. Let
     $s^2 = C\,\mathrm{Tr}(\boldsymbol{\Sigma})$. Estimate $\mathrm{Tr}(\boldsymbol{\Sigma})$ by estimating
     1d variance along $n$ orthogonal directions, see
     Section II-D.
  3) Set $w_i = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{a}\|_2^2}{s^2}\right)$ for every $\boldsymbol{x}_i \in S$.
  4) Return $(S, \boldsymbol{w})$.
```

The second procedure for outlier removal returns a subset of points. It will be convenient to view this as a $0/1$ weighting of the point set. We call this procedure *OutlierTruncation*.

Algorithm 2: OUTLIERTRUNCATION(S, η)

```
Input: $S \subset \mathbb{R}^n, \eta \in [0, 1]$
Output: $\widetilde{S} \subset S, \boldsymbol{w} = \mathbf{1} \in \mathbb{R}^m$
  1) if $n = 1$:
     Let $[t_1, t_2]$ be the smallest interval containing $(1 - \eta - \epsilon)(1 - \eta)$ fraction of the points, $\widetilde{S} \leftarrow S \cap [t_1, t_2]$. Return $(\widetilde{S}, 1)$.
  2) for $i = 1...n$

       a) Let $P_{e_i}\left(\widetilde{S}\right)$ be the projection of $\widetilde{S}$ along
          $i$'th coordinate direction. Let $[t_1, t_2]$ be the
          smallest interval containing $(1-\eta-\epsilon)(1-\eta)$
          fraction of the points.
       b) Let $a_i = \mathrm{mean}\{P_{e_i}\left(\widetilde{S}\right) \leftarrow S \cap [t_1, t_2].\}$
  3) Let $\boldsymbol{a} = (a_1, ..., a_n)$
  4) Let $B(r, \boldsymbol{a})$ = ball of minimum radius $r$ centered
     at $\boldsymbol{a}$ that contains $(1 - \eta - \epsilon)(1 - \eta)$ fraction of
     $S$.
  5) $\widetilde{S} \leftarrow S \cap B(r, \boldsymbol{a})$. Return $(\widetilde{S}, \mathbf{1})$.
```

2) Main Algorithm: We are now ready to state the main algorithm for agnostic mean estimation. It uses one of the above outlier removal procedures and assumes that the output of the procedure is a weighting.

Algorithm 3: AGNOSTICMEAN(S)

```
Input: $S \subset \mathbb{R}^n$, and a routine OUTLIERREMOVAL($\cdot$).
Output: $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^n$.
  1) Let $(\widetilde{S}, \boldsymbol{w})$ = OUTLIERREMOVAL($S$) .
  2) if $n = 1$:
       a) if $\boldsymbol{w} = -1$, Return median($\widetilde{S}$). //Gaussian
          case
       b) else Return mean($\widetilde{S}$). //General case
  3) Let $\boldsymbol{\Sigma}_{\widetilde{S},\boldsymbol{w}}$ be the weighted covariance matrix of
     $\widetilde{S}$ with weights $\boldsymbol{w}$, and $V$ be the span of the top
```

```
   $n/2$ principal components of $\boldsymbol{\Sigma}_{\widetilde{S},\boldsymbol{w}}$, and $W$ be its complement.
4) Set $S_1 := \boldsymbol{P}_V(S)$ where $\boldsymbol{P}_V$ is the projection operation on to $V$.
5) Let $\widehat{\boldsymbol{\mu}}_V :=$ AgnosticMean$(S_1)$ and $\widehat{\boldsymbol{\mu}}_W :=$ mean$(\boldsymbol{P}_W \widetilde{S})$.
6) Let $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^n$ be such that $\boldsymbol{P}_V \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_V$ and $\boldsymbol{P}_W \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_W$.
7) **Return** $\widehat{\boldsymbol{\mu}}$.
```

3) Estimation of the Covariance Matrix and Operator Norm: The algorithm for estimating the covariance matrix calls AgnosticMean on $\boldsymbol{x}\boldsymbol{x}^T$.

```
Algorithm 4: CovarianceEstimation(S)
Input: $S \subset \mathbb{R}^n, \eta \in \mathbb{R}$
Output: $n \times n$ matrix $\widehat{\boldsymbol{\Sigma}}$
  1) $\boldsymbol{x}'_i = \frac{\boldsymbol{x}_i - \boldsymbol{x}_{i+m/2}}{\sqrt{2}}$ for $i \in \{1, ..., |S|/2\}$
  2) Let $S^{(2)} = \{\boldsymbol{x}'_i \boldsymbol{x}'_i \,|\, i = 1, ..., m/2\}$
  3) Run the mean estimation algorithm on $S^{(2)}$, where elements of $S^{(2)}$ are viewed as vectors in $\mathbb{R}^{n^2}$. Let the output be $\widehat{\boldsymbol{\Sigma}}$.
  4) Return $\widehat{\boldsymbol{\Sigma}}$.
```

The algorithm for estimating $\|\boldsymbol{\Sigma}\|_2$, is based on iteratively truncating the samples along the direction of top variance.

```
Algorithm 5: AgnosticOperatorNorm(S)
Input: $S \subset \mathbb{R}^n, \eta \in [0,1], \gamma \in \mathbb{R}$
Output: $\sigma^2 \in \mathbb{R}_{>0}$.
  1) Let $\widetilde{S} =$ SafeOutlierTruncation$(S, \eta, \gamma)$.
  2) Do the following $O(n \log^{2/\gamma} \frac{n}{\eta})$ times
  3) Let $\boldsymbol{\Sigma}_{\mathbf{0}}(\widetilde{S}) := \frac{1}{|\widetilde{S}|} \sum_{i \in \widetilde{S}} \boldsymbol{x}\boldsymbol{x}^T$.
  4) Find $\boldsymbol{v}$, the top eigenvector of $\boldsymbol{\Sigma}_{\mathbf{0}}(\widetilde{S})$, and its corresponding eigenvalue $\sigma^2$.
  5) Estimate up to $1 \pm c\eta$ factor the variance of $\mathcal{D}$ along $\boldsymbol{v}$ and denote it by $\widehat{\sigma}^2_{\boldsymbol{v}}$.
  6) if $\sigma^2 \leq (1 + c_3 \eta \log^{2/\gamma} \frac{n}{\eta}) \widehat{\sigma}^2_{\boldsymbol{v}}$
     Return $\sigma^2$.
  7) Remove all points $\boldsymbol{x} \in \widetilde{S}$ such that $|\boldsymbol{x}^T \boldsymbol{v}| > \frac{c_2 \widehat{\sigma}_v \log^{1/\gamma} \frac{n}{\eta}}{2}$.
  8) Go to Step (3).
```

```
Algorithm 6: SafeOutlierTruncation$(S, \eta, \gamma)$
Input: $S \subset \mathbb{R}^n, \eta \in [0,1], \gamma \in \mathbb{R}$
Output: $\widetilde{S} \subset S$
  1) Let $t = \sum_{i=1}^n \widehat{\sigma}^2_{e_i}$ be the sum of estimated variances of $\mathcal{D}$ in $n$ orthogonal directions.
  2) Let $B(c\sqrt{t} \log^{1/\gamma} \frac{n}{\eta}, \mathbf{0})$ be the ball of radius $c\sqrt{t} \log^{1/\gamma} \frac{n}{\eta}$ centered at $\mathbf{0}$.
  3) $\widetilde{S} \leftarrow S \cap B(c\sqrt{t} \log^{1/\gamma} \frac{n}{\eta}, \mathbf{0})$. Return $\widetilde{S}$.
```

C. Sample Complexity

At various points in the analysis, to bound the sample complexity we will have to show that the estimates computed from samples are close to their expectations. We will use the following two results. Firstly, as an immediate corollary of matrix Bernstien for rectangular matrices (see Theorem 1.6 in [46]), we get the following concentration result for the sample mean and sample covariance.

Lemma II.2. *Consider a distribution in $\mathbb{R}^n$ with covariance matrix $\boldsymbol{\Sigma}$ and supported in some Euclidean ball whose radius we denote is $\sqrt{R\|\boldsymbol{\Sigma}\|}$, for some $R \in \mathbb{R}$. Let $\epsilon \in (0,1)$. Then the following holds with probability at least $1 - 1/\operatorname{poly}(n)$: If $N \geq \frac{R \log n}{\epsilon^2}$ then*

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \epsilon \sqrt{\|\boldsymbol{\Sigma}\|}$$

and

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leq \epsilon \|\boldsymbol{\Sigma}\|.$$

Here $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are sample mean and sample covariance matrix.

Secondly, the functions we estimate will be integrals of low-degree polynomials (degree d at most 4) restricted to intervals and/or balls. These functions viewed as binary concepts have small VC-dimension, $O(n^d)$ where n is the dimension of space and d is the degree of the polynomial. We use this to bound the error of estimating integrals via samples, and we can make the error smaller than any inverse polynomial using a $\operatorname{poly}(n)$ size sample.

Proposition II.3. *Let F be a class of real-valued functions from $\mathbb{R}^n$ to $[-R, R]$. Let C_F be the corresponding class of binary concepts, i.e., for each $f \in F$, we consider the concepts $h_t(x) = 1$ if $f(x) \geq t$ and zero otherwise. Suppose the VC-dimension of C_F is d. Then, for any $f \in F$, and any distribution $\mathcal{D}$ over $\mathbb{R}^n$, an iid sample S of size $|S| \geq \frac{8}{\epsilon^2}(d \log(1/\epsilon) + \log(1/\delta))$, with probability at least $1 - \delta$ satisfies*

$$\left| \mathbf{E}_{x \sim \mathcal{D}}(f(x)) - \frac{1}{|S|} \sum_{x \in S} f(x) \right| \leq 2\epsilon R.$$

Proof: By the VC theorem, for any concept in C_F, the bound on the size of the sample ensures that with probability at least $1-\delta$ and any t,

$$\left|\Pr(f(x)\geq t)-\frac{|\{x\in S\,:\,f(x)\geq t\}|}{|S|}\right|\leq\epsilon.$$

Noting that $\mathbf{E}_{x\sim\mathcal{D}}(f(x))=\int_{-R}^{R}\Pr(f(x)\geq t)\,dt$, we get the claimed bound.

□

D. 1d Estimation

In this section, we give a brief descriptions of the algorithm that we can use to find mean and the variance in one dimension. We will first consider the case when $\mathcal{D}=N(\mu,\sigma^2)$. Suppose we are given samples $S=\{x_1,...,x_m\}$ from $\mathcal{D}_\eta$, and furthermore that $\eta<1/2.1$ (a constant strictly less than 1/2 suffices). We can then estimate the mean by the median x_{median} of S. We can prove the following statement

Lemma II.4. *Let $\mathcal{D}=N(0,\sigma^2)$ be a one dimensional Gaussian distribution. If $m=O\left(\frac{\log n}{\epsilon^2}\right)$, and we are given $x_1,...,x_m\sim\mathcal{D}_\eta$. With probability $1-1/\operatorname{poly}(n)$ the following hold.*
Mean: *The median $x_{med}=\operatorname{median}_i\{x_i\}$ satisfies $|x_{\text{median}}|=O(\eta+\epsilon)\sigma$ with probability $1-1/\operatorname{poly}(n)$.*
Variance: *There is an algorithm that computes in polynomial time $\widehat{\sigma}^2$ such that $\left|\widehat{\sigma}^2-\sigma^2\right|=O(\eta+\epsilon)\sigma^2$.*

The proof follows from Hoeffding's inequality and the fact that $\Phi^{-1}(1/2+\eta+\epsilon)=O(\eta+\epsilon)\sigma$ when $\eta+\epsilon<1/2.05$. Here Φ is the c.d.f. of a standard normal variable. We use a very similar idea to estimate the variance in this case. We look at a another quantile of S in addition to the median and the estimate is obtained by using both the quantiles.

In the case when $\mathcal{D}$ just has bounded fourth moments, median cannot be used as an estimate. In fact, there are distributions (Bernoulli) for which median does poorly even when there is no noise. Therefore, we need a different method in this case. Suppose $|S|=\Omega\left(\frac{\log n+\log 1/\epsilon}{\epsilon^2}\right)$. We consider the interval of minimum length that contains $(1-\eta-\epsilon)(1-\eta)$ fraction of the sample points. Our estimator $\widehat{\mu}$ for the mean then is sample mean of all the points that lie in this interval. To estimate the variance we compute the sample variance of points in this interval. We can show the following guarantee for mean estimation

Lemma II.5. *If $\boldsymbol{x}\sim\mathcal{D}$ has bounded fourth moments with constant C_4, and $(x-\mu)^2$ has bounded fourth moments with constant $C_{4,2}$. Let S be samples from $\mathcal{D}_\eta$ such that $|S|=\Omega\left(\frac{\log n+\log 1/\epsilon}{\epsilon^2}\right)$. With probability $1-1/\operatorname{poly}(n)$ the following hold.*
Mean: *The mean estimator as defined above satisfies $|\widehat{\mu}-\mu|=O\left(C_4^{1/4}(\eta+\epsilon)^{3/4}\right)\sigma$.*
Variance: *Variance estimator $\widehat{\sigma}^2$ as defined above satisfies $\left|\widehat{\sigma}^2-\sigma^2\right|=O\left(C_{4,2}^{1/4}(\eta+\epsilon)^{3/4}C_4^{1/2}\sigma\right)$.*

E. Proof of the Main Theorem

Here we give an outline of the proof of the main theorem. Since the high level structure for both the normal distribution and distribution with bounded fourth moments are same, we will focus on the normal case. We therefore assume $\mathcal{D}=N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ in this section. Let $s^2:=\frac{1}{\epsilon_1}\operatorname{Tr}(\boldsymbol{\Sigma})$ and $\epsilon_2:=\frac{\|\boldsymbol{a}\|_2^2}{\eta^2s^2}$. We can estimate $\operatorname{Tr}(\boldsymbol{\Sigma})$ by estimating (1 dimensional) variances along n orthogonal directions, see Section II-D. Note that we can arrange $0<\epsilon_1,\epsilon_2<1$ to be small enough constants. Let $\boldsymbol{a}$ be the coordinate-wise median, we can show $\|\boldsymbol{a}\|_2^2\leq C\eta^2\operatorname{Tr}(\boldsymbol{\Sigma})$ with probability $1-1/\operatorname{poly}(n)$. We weight every point $\boldsymbol{x}$ by $w_{\boldsymbol{x}}=\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{a}\|^2}{s^2})$. Let $S=\{\boldsymbol{x}_1,...,\boldsymbol{x}_m\},\boldsymbol{x}_i\sim\mathcal{D}_\eta$ be the sample we get. Let $S=S_G\cup S_N$ be the Gaussian and the noise points repectively, with $|S_N|=\eta m$. For a set $T\subset\mathbb{R}^n$, let

$$\boldsymbol{\mu}_{T,\boldsymbol{w}}:=\frac{1}{m}\sum_{i\in T}w_{\boldsymbol{x}_i}\boldsymbol{x}_i\ \text{ and}$$
$$\boldsymbol{\Sigma}_{T,\boldsymbol{w}}:=\frac{1}{|T|}\sum_{i\in T}w_i(\boldsymbol{x}_i-\boldsymbol{\mu}_{T,\boldsymbol{w}})(\boldsymbol{x}_i-\boldsymbol{\mu}_{T,\boldsymbol{w}})^T$$

We use the above notation for $T=S_G$ and $T=S_N$. By an abuse of notation, when $T=G$, we mean the population version of the above quantities:

$$\boldsymbol{\mu}_{G,\boldsymbol{w}}:=\mathbf{E}_{\boldsymbol{x}}w_{\boldsymbol{x}}\boldsymbol{x}\ \text{ and }\ \boldsymbol{\Sigma}_{G,\boldsymbol{w}}:=\mathbf{E}_{\boldsymbol{x}}w_{\boldsymbol{x}}(\boldsymbol{x}-\boldsymbol{\mu}_{G,\boldsymbol{w}})(\boldsymbol{x}_i-\boldsymbol{\mu}_{G,\boldsymbol{w}})^T.$$

Note that

$$\boldsymbol{\mu}_{S,\boldsymbol{w}}=(1-\eta)\boldsymbol{\mu}_{S_G,\boldsymbol{w}}+\eta\boldsymbol{\mu}_{S_N,\boldsymbol{w}}.$$

We consider the matrix $\boldsymbol{\Sigma}_{S,\boldsymbol{w}}$

$$\begin{aligned}\boldsymbol{\Sigma}_{S,\boldsymbol{w}}&=\frac{1}{m}\sum_i w_{\boldsymbol{x}_i}(\boldsymbol{x}_i-\boldsymbol{\mu}_{S,\boldsymbol{w}})(\boldsymbol{x}_i-\boldsymbol{\mu}_{S,\boldsymbol{w}})^T\\&=(1-\eta)\boldsymbol{\Sigma}_{S_G,\boldsymbol{w}}+\eta\boldsymbol{\Sigma}_{S_N,\boldsymbol{w}}\\&\quad+\eta(1-\eta)(\boldsymbol{\mu}_{S_N,\boldsymbol{w}}-\boldsymbol{\mu}_{S_G,\boldsymbol{w}})(\boldsymbol{\mu}_{S_N,\boldsymbol{w}}-\boldsymbol{\mu}_{S_G,\boldsymbol{w}})^T.\end{aligned}$$

We first show that the covariance matrix doesn't change much because of outlier damping. For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, by $\boldsymbol{A}\preceq\boldsymbol{B}$ we mean that $\boldsymbol{B}-\boldsymbol{A}$ is a positive semidefinite matrix.

Lemma II.6. *We have*

$$\left(\frac{e^{-\eta^2\epsilon_2}}{1+\epsilon_1}-\eta^2\epsilon_2e^{2\epsilon_1}\right)\boldsymbol{\Sigma}\preceq\boldsymbol{\Sigma}_{G,\boldsymbol{w}}\preceq e^{\epsilon_1}\boldsymbol{\Sigma}$$

and $\|\boldsymbol{\mu}_{G,w}-\boldsymbol{\mu}\|=O\left(\eta\sqrt{\|\boldsymbol{\Sigma}\|}\right)$.

It is important to note that when $\boldsymbol{\Sigma}=\sigma^2\boldsymbol{I}$ is a multiple of identity, then $\boldsymbol{\Sigma}_{G,\boldsymbol{w}}$ will also be a multiple of $\boldsymbol{I}$. By Lemma II.2, if we take $m=O(\frac{n\log n}{\epsilon^2})$ samples, we will have

$$(1-\epsilon)\boldsymbol{\Sigma}_{G,\boldsymbol{w}}\preceq\boldsymbol{\Sigma}_{S_G,\boldsymbol{w}}\preceq(1+\epsilon)\boldsymbol{\Sigma}_{G,\boldsymbol{w}}.\tag{6}$$

The next step is to give guarantees for one level of the projection (one iteration of Algorithm 3). Suppose, we have

$$\alpha\boldsymbol{\Sigma} \preceq \boldsymbol{\Sigma}_{S_G,\boldsymbol{w}} \preceq \beta\boldsymbol{\Sigma}$$

for some $\alpha, \beta > 0$. By an argument similar to the one sketched in Section II, we can prove

Lemma II.7. *We will use the notation as defined above. Let W be the bottom $n/2$ principal components of the covariance matrix $\boldsymbol{\Sigma}_{S,\boldsymbol{w}}$. We have*

$$\|\eta P_W \boldsymbol{\delta}_{\boldsymbol{\mu}}\|^2 \le 2\eta\left((\beta + C\eta)\|\boldsymbol{\Sigma}\|_2 - \alpha\|\boldsymbol{\Sigma}\|_{\min}\right),$$

where $\|\boldsymbol{\Sigma}\|_{\min}$ denotes the least eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\delta}_{\boldsymbol{\mu}} := \boldsymbol{\mu}_{S_N,\boldsymbol{w}} - \boldsymbol{\mu}_{S_G,\boldsymbol{w}}$.

By an inductive application of Lemma II.7, we get the following theorem giving a bound on $\|\widehat{\boldsymbol{\mu}}\|$.

Theorem II.8. *On input S and the routine* OUTLIERDAMPING$(\cdot)$*,* AGNOSTICMEAN *outputs $\widehat{\boldsymbol{\mu}}$ satisfying*

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \le O\left((\beta\eta + \eta^2 + \epsilon^2)\|\boldsymbol{\Sigma}\|_2 - \alpha\eta\|\boldsymbol{\Sigma}\|_{\min}\right)(1+\log n).$$

Lemma II.6 combined with Equation 6 and Theorem II.8 proves Theorem I.1. We get a better dependence on η when $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$ because we can take $\alpha = \beta$ in this case. This would lead to the cancellation of the leading term in the bound in Theorem II.8 as $\|\boldsymbol{\Sigma}\|_2 = \|\boldsymbol{\Sigma}\|_{\min}$.

OPEN QUESTIONS

An immediate open question is whether the our analysis of the mean estimation algorithm is tight and the $\sqrt{\log n}$ is avoidable. For special distributions including Gaussians, [44] give an algorithm with higher sample complexity and error $\eta\sqrt{\log\frac{1}{\eta}}$ rather than $\eta\sqrt{\log n}$ or $\sqrt{\eta \log n}$ as in Theorem I.1. An open question is to give an $O(\eta)$ approximation. For the more general distributions considered here, the dependence on η must grow as at least $\eta^{3/4}$; it is open to find an algorithm that achieves $O(\eta^{3/4})$ error (our guarantee for the general setting has error $O(\sqrt{\eta \log n})$). Other open problems include agnostic learning of a mixture of two arbitrary Gaussians and agnostic sparse recovery.

ACKNOWLEDGMENT

We thank Chao Gao and Roman Vershynin for helpful discussions. We would also like to thank the anonymous reviewers for useful suggestions. This research was supported in part by NSF awards CCF-1217793 and EAGER-1555447.

REFERENCES

[1] A. T. Kalai, A. Moitra, and G. Valiant, “Efficiently learning mixtures of two Gaussians,” in *Proceedings of the 42nd ACM symposium on Theory of computing*. ACM, 2010, pp. 553–562.

[2] A. Moitra and G. Valiant, “Settling the polynomial learnability of mixtures of Gaussians,” in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, 2010, pp. 93–102.

[3] M. Belkin and K. Sinha, “Polynomial learning of distribution families,” in *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, 2010, pp. 103–112. [Online]. Available: http://dx.doi.org/10.1109/FOCS.2010.16

[4] S. Dasgupta, “Learning mixtures of Gaussians,” in *Foundations of Computer Science, 1999. 40th Annual Symposium on*. IEEE, 1999, pp. 634–644.

[5] S. Arora and R. Kannan, “Learning mixtures of arbitrary Gaussians,” in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*. ACM, 2001, pp. 247–257.

[6] S. Vempala and G. Wang, “A spectral algorithm for learning mixture models,” *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 841–860, 2004.

[7] S. Dasgupta and L. Schulman, “A probabilistic analysis of EM for mixtures of separated, spherical Gaussians,” *The Journal of Machine Learning Research*, vol. 8, pp. 203–226, 2007.

[8] K. Chaudhuri and S. Rao, “Learning mixtures of product distributions using correlations and independence,” in *Proc. of COLT*, 2008.

[9] S. C. Brubaker and S. S. Vempala, “Isotropic PCA and affine-invariant clustering,” in *Building Bridges*. Springer, 2008, pp. 241–281.

[10] D. Hsu and S. M. Kakade, “Learning mixtures of spherical Gaussians: moment methods and spectral decompositions,” in *ITCS*, 2013, pp. 11–20.

[11] A. M. Frieze, M. Jerrum, and R. Kannan, “Learning linear transformations,” in *FOCS*, 1996, pp. 359–368.

[12] P. Q. Nguyen and O. Regev, “Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures,” *J. Cryptology*, vol. 22, no. 2, pp. 139–160, 2009.

[13] J.-F. Cardoso, “Multidimensional independent component analysis,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 4. IEEE, 1998, pp. 1941–1944.

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[15] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation*. Academic Press, 2010.

[16] M. Belkin, L. Rademacher, and J. Voss, “Blind signal separation in the presence of Gaussian noise,” in *Proc. of COLT*, 2013.

[17] S. Arora, R. Ge, A. Moitra, and S. Sachdeva, “Provable ICA with unknown gaussian noise, with implications for gaussian mixtures and autoencoders,” in *NIPS*, 2012, pp. 2384–2392.

[18] A. Bhaskara, M. Charikar, and A. Vijayaraghavan, “Uniqueness of tensor decompositions with applications to polynomial identifiability,” *arXiv preprint arXiv:1304.8087*, 2013.

[19] N. Goyal, S. Vempala, and Y. Xiao, “Fourier pca and robust tensor decomposition,” in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM, 2014, pp. 584–593.

[20] S. Vempala and Y. Xiao, “Max vs min: Tensor decomposition and ICA with nearly linear sample complexity,” in *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, 2015, pp. 1710–1723. [Online]. Available: http://jmlr.org/proceedings/papers/v40/Vempala15.html

[21] R. Kannan and S. Vempala, *Spectral Algorithms*. Now Publishers Inc, 2009.

[22] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition,” *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[23] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, Jun. 2011. [Online]. Available: http://doi.acm.org/10.1145/1970392.1970395

[24] H. Xu, C. Caramanis, and S. Mannor, “Principal component analysis with contaminated data: The high dimensional case,” *arXiv preprint arXiv:1002.4658*, 2010.

[25] N. Kwak, “Principal component analysis based on l1-norm maximization,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 9, pp. 1672–1680, 2008.

[26] M. McCoy, J. A. Tropp *et al.*, “Two proposals for robust pca using semidefinite programming,” *Electronic Journal of Statistics*, vol. 5, pp. 1123–1160, 2011.

[27] P. J. Huber, *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, ch. Robust Statistics, pp. 1248–1251. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04898-2_594

[28] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011, vol. 114.

[29] R. Maronna, D. Martin, and V. Yohai, *Robust statistics*. John Wiley & Sons, Chichester. ISBN, 2006.

[30] P. J. Huber, “Robust estimation of a location parameter,” *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03 1964. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177703732

[31] J. W. Tukey, “Mathematics and the Picturing of Data,” in *International Congress of Mathematicians 1974*, R. D. James, Ed., vol. 2, 1974, pp. 523–532.

[32] R. A. Maronna, “Robust m-estimators of multivariate location and scatter,” *Ann. Statist.*, vol. 4, no. 1, pp. 51–67, 01 1976. [Online]. Available: http://dx.doi.org/10.1214/aos/1176343347

[33] J. R. K. S. J. Devlin, R. Gnandesikan, “Robust estimation of dispersion matrices and principal components,” *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 354–362, 1981. [Online]. Available: http://www.jstor.org/stable/2287836

[34] D. L. Donoho, “Breakdown properties of multivariate location estimators,” Ph.D. dissertation, Harvard University, 1982.

[35] P. L. Davies, “Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices,” *Ann. Statist.*, vol. 15, no. 3, pp. 1269–1292, 09 1987. [Online]. Available: http://dx.doi.org/10.1214/aos/1176350505

[36] P. J. R. Hendrik P. Lopuhaa, “Breakdown points of affine equivariant estimators of multivariate location and covariance matrices,” *The Annals of Statistics*, vol. 19, no. 1, pp. 229–248, 1991. [Online]. Available: http://www.jstor.org/stable/2241852

[37] D. L. Donoho and M. Gasko, “Breakdown properties of location estimates based on halfspace depth and projected outlyingness,” *Ann. Statist.*, vol. 20, no. 4, pp. 1803–1827, 12 1992. [Online]. Available: http://dx.doi.org/10.1214/aos/1176348890

[38] R. A. Maronna, W. A. Stahel, and V. J. Yohai, “Bias-robust estimators of multivariate scatter based on projections,” *J. Multivar. Anal.*, vol. 42, no. 1, pp. 141–161, Jul. 1992. [Online]. Available: http://dx.doi.org/10.1016/0047-259X(92)90084-S

[39] R. A. Maronna and R. H. Zamar, “Robust estimates of location and dispersion for high-dimensional datasets,” *Technometrics*, 2012.

[40] M. Chen, C. Gao, and Z. Ren, “Robust Covariance Matrix Estimation via Matrix Depth,” *ArXiv e-prints*, Jun. 2015.

[41] C. G. Small, “A survey of multidimensional medians,” *International Statistical Review/Revue Internationale de Statistique*, pp. 263–277, 1990.

[42] D. Bruce, “A multivariate median in banach spaces and applications to robust pca,” http://www-personal.umich.edu/ romanv/students/bruce-REU.pdf, 2011.

[43] https://github.com/kal2000/AgnosticMeanAndCovarianceCode.

[44] I. Diakonikolas, G. Kamath, D. M. Kane, J. Z. Li, A. Moitra, and A. Stewart, “Robust estimators in high dimensions without the computational intractability,” *CoRR*, vol. abs/1604.06443, 2016. [Online]. Available: http://arxiv.org/abs/1604.06443

[45] S. C. Brubaker, “Robust PCA and clustering in noisy mixtures,” in *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4-6, 2009*, 2009, pp. 1078–1087. [Online]. Available: http://dl.acm.org/citation.cfm?id=1496770.1496887

[46] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.