Robust Moment Estimation and Improved Clustering via Sum of Squares*

Pravesh K. Kothari
Princeton University and the Institute for Advanced Study
Princeton, NJ, USA

Jacob Steinhardt
Stanford University
Palo Alto, CA, USA

David Steurer
ETH Zurich
Zurich, Switzerland

ABSTRACT

We develop efficient algorithms for estimating low-degree moments of unknown distributions in the presence of adversarial outliers and design a new family of convex relaxations for k-means clustering based on sum-of-squares method. As an immediate corollary, for any $\gamma > 0$, we obtain an efficient algorithm for learning the means of a mixture of k arbitrary Poincaré distributions in $\mathbb{R}^d$ in time $d^{O(1/\gamma)}$ so long as the means have separation $\Omega(k^\gamma)$. This in particular yields an algorithm for learning Gaussian mixtures with separation $\Omega(k^\gamma)$, thus partially resolving an open problem of Regev and Vijayaraghavan (2017). The guarantees of our robust estimation algorithms improve in many cases significantly over the best previous ones, obtained in the recent works. We also show that the guarantees of our algorithms match information-theoretic lower-bounds for the class of distributions we consider. These improved guarantees allow us to give improved algorithms for independent component analysis and learning mixtures of Gaussians in the presence of outliers.

We also show a sharp upper bound on the sum-of-squares norms for moment tensors of any distribution that satisfies the *Poincaré inequality*. The Poincaré inequality is a central inequality in probability theory, and a large class of distributions satisfy it including Gaussians, product distributions, strongly log-concave distributions, and any sum or uniformly continuous transformation of such distributions. As a consequence, this yields that all of the above algorithmic improvements hold for distributions satisfying the Poincaré inequality.

CCS CONCEPTS

• **Theory of computation → Unsupervised learning and clustering**;

*The conference version is a union of two papers available here: https://arxiv.org/abs/1711.11581 and https://arxiv.org/abs/1711.07465

STOC'18, June 25–29, 2018, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5559-9/18/06…$15.00
https://doi.org/10.1145/3188745.3188970

KEYWORDS

clustering, sum-of-squares, robust learning, moments

ACM Reference Format:
Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. 2018. Robust Moment Estimation and Improved Clustering via Sum of Squares. In *Proceedings of 50th Annual ACM SIGACT Symposium on the Theory of Computing (STOC'18).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3188745.3188970

1 INTRODUCTION

Progress on many fundamental unsupervised learning tasks has required circumventing a plethora of intractability results by coming up with natural restrictions on input instances that preserve some essential character of the problem. For example, while k-means clustering is NP-hard in the worst-case [51], there is an influential line of work providing spectral algorithms for clustering mixture models satisfying appropriate assumptions [1, 7, 45]. On the flip side, we run the risk of developing algorithmic strategies that exploit strong assumptions in a way that makes them brittle. We are thus forced to walk the tight rope of avoiding computational intractability without "overfiting" our algorithmic strategies to idealized assumptions on input data.

Consider, for example, the problem of clustering data into k groups. On the one hand, a line of work leading to [7] shows that a variant of spectral clustering can recover the underlying clustering so long as each cluster has bounded covariance around its center and the cluster centers are separated by at least $\Omega(\sqrt{k})$. Known results can improve on this bound to require a separation of $\Omega(k^{1/4})$ if the cluster distributions are assumed to be isotropic and log-concave [63]. If the cluster means are in general position, other lines of work yields results for Gaussians [6, 12, 14, 16, 28, 31, 36, 41, 53] or for distributions satisfying independence assumptions [4, 38]. However, the assumptions often play a crucial role in the algorithm. For example, the famous method of moments that yields a result for learning mixtures of Gaussians in general position uses the specific algebraic structure of the moment tensor of Gaussian distributions. Such techniques are unlikely to work for more general classes of distributions.

As another example, consider the robust mean estimation problem which has been actively investigated recently. Lai et. al. [46] and later improvements [24, 61] show how to estimate the mean of an unknown distribution (with bounded second moments) where an

ε fraction of points are adversarially corrupted, obtaining additive error $O(\sqrt{\varepsilon})$. On the other hand, Diakonikolas et. al. [23] showed how to estimate the mean of a Gaussian or product distribution with nearly optimal additive error $\tilde{O}(\varepsilon)$. However, their algorithm again makes strong use of the known algebraic structure of the moments of these distributions.

Further scrutiny reveals that the two examples of clustering and robust mean estimation suffer from a "second-moment" barrier. For both problems, the most general results algorithmically exploit only some boundedness condition on the second moments of the data, while the strongest results use *exact* information about higher moments (e.g. by assuming Gaussianity) and are thus brittle. This leads to the key conceptual driving force of the present work:

Can we algorithmically exploit boundedness information about a limited number of low-degree moments?

As the above examples illustrate, this is a natural way to formulate the "in-between" case between the two well-explored extremes. From an algorithmic perspective, this question forces us to develop techniques that can utilize information about higher moments of data for problems such as clustering and mean estimation. For these problems, we can more concretely ask:

Can we beat the second-moment barrier in the agnostic setting for clustering and robust mean estimation?

The term *agnostic* here refers to the fact that we want our algorithm to work for as wide a class of distributions as possible, and in particular to avoid making parametric assumptions (such as Gaussianity) about the underlying distribution.

The main goal of this work is to present a principled way to utilize higher moment information in input data and break the second moment barrier for both clustering and robust estimation of basic parameters such as mean, covariance and in general, higher moments of distributions.

Outlier-Robust Parameter Estimation. We consider the problem of *outlier-robust parameter estimation*: We are given independent draws $x_1, \ldots, x_n$ from an unknown distribution D over $\mathbb{R}^d$ and the goal is to estimate parameters of the distribution, e.g., its mean or its covariance matrix. Furthermore, we require that the estimator is *outlier-robust*: even if an ε-fraction of the draws are corrupted by an adversary, the estimation error should be small.

These kind of estimators have been studied extensively in statistics (under the term *robust statistics*) [33, 40, 52, 62]. However, many robust estimators coming out of this research effort are computationally efficient only for low-dimensional distributions (because the running time is say exponential in the dimension d) [13].

A recent line of research developed the first robust estimators for basic parameter estimation problems (e.g., estimating the mean and covariance matrix of a Gaussian distribution) that are computationally efficient for the high-dimensional case, i.e., the running time is only polynomial in the dimension d [19, 21, 25, 47].

Our work continues this line of research. We design efficient algorithms to estimate low-degree moments of distributions. Our estimators succeed under significantly weaker assumptions about the unknown distribution D, even for the most basic tasks of estimating the mean and covariance matrix of D. For example, in order to estimate the mean of D (in an appropriate norm) our algorithms do not need to assume that D is Gaussian or has a covariance matrix with small spectral norm in contrast to assumptions of previous works. Similary, our algorithms for estimating covariance matrices work, unlike previous algorithms, for non-Gaussian distributions and distributions that are not (affine transformations of) product distributions.

Besides these qualitative differences, our algorithms also offer quantitative improvements. In particular, for the class of distributions we consider, the guarantees of our algorithms—concretely, the asymptotic behavior of the estimation error as a function of the fraction ε of corruptions—match, for the first time in this generality, information-theoretic lower bounds.

Outlier-robust method of moments. Our techniques for robust estimation of mean vectors and covariance matrices extend in a natural way to higher-order moment tensors. This fact allows us to turn many non-outlier-robust algorithms in a black-box way into outlier-robust algorithms. The reason is that for many parameter estimation problems the best known algorithms in terms of provable guarantees are based on the *method of moments*, which means that they don't require direct access to a sample from the distribution but instead only to its low-degree moments (e.g. [42, 54, 56]). (Often, a key ingredient of these algorithms in the high-dimensional setting is *tensor decomposition* [3, 5, 10, 15, 29, 39, 49, 55].) If there were no outliers, we could run these kinds of algorithms on the empirical moments of the observed sample. However, in the presence of outliers, this approach fails dramatically because even a single outlier can have a huge effect on the empirical moments. Instead, we apply method-of-moment-based algorithms on the output of our outlier-robust moment estimators. Following this strategy, we obtain new outlier-robust algorithms for independent component analysis (even if the underlying unknown linear transformation is ill-conditioned) and mixtures of spherical Gaussians (even if the number of components of the mixture is large and the means are not separated).

Estimation algorithms from identifiability proofs. Our algorithms for outlier-robust parameter estimation and their analysis follow a recent paradigm for computationally-efficient provable parameter estimation that has been developed in the context of the *sum-of-squares method*. We say that a parameter estimation problem satisfies *identifiability* if it is information-theoretically possible to recover the desired parameter from the available data (disregarding computational efficiency). The key idea of this paradigm is that a proof of identifiability can be turned into an efficient estimation algorithm if the proof is captured by a low-complexity proof system

like sum-of-squares. Many estimation algorithms based on convex relaxations, in particular sum-of-squares relaxations, can be viewed as following this paradigm (e.g., compressed sensing and matrix completion [17, 18, 32, 58]). Moreover, this paradigm has been used, with a varying degree of explicitness, in order to design a number of recent algorithms based on sum-of-squares for unsupervised learning, inverse, and estimation problems like overcomplete tensor decomposition, sparse dictionary learning, tensor completion, tensor principal component analysis [10, 11, 35, 49, 57].

In many of these settings, including ours, the proof of identifiability takes a particular form: given a (corrupted) sample $X \subseteq \mathbb{R}^d$ from a distribution D and a candidate parameter $\hat{\theta}$ that is close to true parameter θ of D, we want to be able to efficiently certify that the candidate parameter $\hat{\theta}$ is indeed close to θ. (This notion of certificate is similar to the one for NP.) Following the above paradigm, if the certification in this identifiability proof can be accomplished by a low-degree sum-of-squares proof, we can derive an efficient estimation algorithm (that computes an estimate $\hat{\theta}$ just given the sample X).

Next we describe the form of our identifiability proofs for outlier-robust estimation.

Identifiability in the presence of outliers. Suppose we are given an ε-corrupted sample $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ of an unknown distribution D and our goal is to estimate the mean $\mu \in \mathbb{R}^d$ of D. To hope for the mean to be identifiable, we need to make some assumptions about D. Otherwise, D could for example have probability mass $1-\varepsilon$ on 0 and probability mass ε on μ/ε. Then, an adversary can erase all information about the mean μ in an ε-corrupted sample from D (because only an ε fraction of the draws from D carry information about μ). Therefore, we will assume that D belongs to some class of distributions C (known to the algorithm). Furthermore, we will assume that the class C is defined by conditions on low-degree moments so that if we take a large enough (non-corrupted) sample X^0 from D, the uniform distribution over X^0 also belongs to the class C with high probability. (We describe the conditions that define C in a paragraph below.)

With this setup in place, we can describe our robust identifiability proof: it consists of a (multi-)set of vectors $X' = \{x'_1, \ldots, x'_n\} \subseteq \mathbb{R}^d$ that satisfies two conditions:

(1) $x'_i = x_i$ for all but an ε fraction of the indices $i \in [n]$,
(2) the uniform distribution over X' is in C.

Note that given X and X', we can efficiently check the above conditions, assuming that the conditions on the low-degree moments that define C are efficiently checkable (which they will be).

Also note that the above notion of proof is complete in the following sense: if X is indeed an ε-corruption of a typical[1] sample X^0 from a distribution $D \in C$, then there exists a set X' that satisfies the above conditions, namely the uncorrupted sample X^0.

We show that the above notion of proof is also sound: if X is indeed an ε-corruption of a (typical) sample X^0 from a distribution $D \in C$ and X' satisfies the above conditions, then the empirical mean $\mu' := \frac{1}{n}\sum_{i=1}^n x'_i$ of the uniform distribution over X' is close to μ. We can rephrase this soundness as the following concise mathematical statement, which we prove in the full version.if D and D' are two distributions in C that have small statistical distance, then their means are close to each other (and their higher-order moments are close to each other as well).

Furthermore, the above soundness is captured by a low-degree sum-of-squares proof (using for example the sum-of-squares version of Hölder's inequality); this fact is the basis of our efficient algorithm for outlier-robust mean estimation.

Outlier-robustness and (certifiable) subgaussianity. To motivate the aforementioned conditions we impose on the distributions, consider the following scenario: we are given an ε-corrupted sample of a distribution D over $\mathbb{R}$ and our goal is to estimate its variance σ^2 up to a constant factors. To hope for the variance to be identifiable, we need to rule out for example that D outputs 0 with probability $1-\varepsilon$ and a Gaussian $N(0, \sigma^2/\varepsilon^2)$ with probability ε. Otherwise, an adversary can remove all information about σ by changing an ε fraction of the draws in a sample from D. If we look at the low-degree moments of this distribution D, we see that $(\mathbb{E}_{D(x)} x^k)^{1/k} \approx \sqrt{k/\varepsilon^{1-2/k}}(\mathbb{E}_{D(x)} x^2)^{1/2}$. So for large enough k (i.e., $k \approx \log(1/\varepsilon)$), the ratio between the L_k norm of D and the L_2 norm of D exceeds that of a Gaussian by a factor of roughly $1/\sqrt{\varepsilon}$. In order to rule out this example, we impose the following condition on the low-degree moments of D, and we show that this condition is enough to estimate the variance of a distribution D over $\mathbb{R}$ up to constant factors given an ε-corrupted sample,

$$\Big(\mathop{\mathbb{E}}_{D(x)} (x-\mu_D)^k\Big)^{1/k} \leqslant \sqrt{Ck} \cdot \Big(\mathop{\mathbb{E}}_{D(x)} (x-\mu_D)^2\Big)^{1/2} \quad \text{for } C > 0 \text{ and even } k \in \mathbb{N} \text{ with } Ck \cdot \varepsilon^{1-2/k} \ll 1\,. \tag{1.1}$$

Here, μ_D is the mean of the distribution D.

In the high-dimensional setting, a natural idea is to impose the condition Eq. (1.1) for every direction $u \in \mathbb{R}^d$,

$$\Big(\mathop{\mathbb{E}}_{D(x)} \langle x-\mu_D, u\rangle^k\Big)^{1/k} \leqslant \sqrt{Ck} \cdot \Big(\mathop{\mathbb{E}}_{D(x)} \langle x-\mu_D, u\rangle^2\Big)^{1/2}\,. \tag{1.2}$$

We show that this condition is indeed enough to ensure identifiability and that it is information-theoretically possible to estimate the covariance matrix Σ_D of D up to constant factors (in the Löwner order sense) assuming again that $Ck \cdot \varepsilon^{1-2/k} \ll 1$. Unfortunately, condition Eq. (1.2) is unlikely to be enough to guarantee an efficient estimation algorithm. The reason is that Eq. (1.2) might hold for the low-degree moments of some distribution D but every proof of this fact requires exponential size. (This phenomenon is related

[1] The sample X^0 of D should be typical in the sense that the empirical low-degree moments of the sample X^0 are close to the (population) low-degree moments of the distribution D. If the sample is large enough (polynomial in the dimension), this condition is satisfied with high probability.

to the fact that finding a vector $u \in \mathbb{R}^d$ that violates Eq. (1.2) is an NP-hard problem in general.)

To overcome this source of intractability, we require that inequality Eq. (1.2) is not only true but also has a low-degree sum-of-squares proof.

Definition 1.1 (Certifiable subgaussianity of low-degree moments)**.** A distribution[2] D over $\mathbb{R}^d$ with mean μ is (k, ℓ)-*certifiably subgaussian* with parameter $C > 0$ if there for every positive integer $k' \leqslant k/2$, there exists a degree-ℓ sum-of-squares proof[3] of the degree-$2k'$ polynomial inequality over the unit sphere,

$$\forall u \in \mathbb{S}^{d-1}. \ \mathop{\mathbb{E}}_{D(x)} \langle x-\mu, u\rangle^{2k'} \leqslant \left(C \cdot k' \mathop{\mathbb{E}}_{D(x)} \langle x-\mu, u\rangle^2\right)^{k'} \quad (1.3)$$

With this additional condition, we give a polynomial-time algorithm to estimate the covariance matrix Σ_D of D up to constant factors (in the Löwner order sense) assuming again that $Ck \cdot \varepsilon^{1-2/k} \ll 1$.

Since any valid polynomial inequality over the sphere has a sum-of-squares proof if we allow the degree to be large enough and the inequality has positive slack, we say that D is (k, ∞)-*certifiably subgaussian* with parameter $C > 0$ if the inequalities Eq. (1.3) hold for all $k' \leqslant k/2$. In most cases we consider, a sum-of-squares proof of degree $\ell = k$ is enough. In this case, we just say that D is k-certifiably subgaussian. Observe that k-certifiable subgaussianity only restricts moments up to degree k of the underlying distribution. In this sense, it is a much weaker assumption than the usual notion of subgaussianity, which imposes Gaussian-like upper bounds on all moments.

Certain certifiably subgaussian distributions have been well-known and utilized in previous results on applications of the SoS method in machine learning. We also discuss some basic examples of certifiably subgaussian distributions such as uniform distribution on the hypercube $\{\pm 1\}^d$ and the any d dimensional gaussian distribution. We also observe that many operations on distributions preserve this property. In particular, (affine transformations of) products of scalar-valued subgaussian distributions and mixtures thereof satisfy this property.

Sum-of-squares and quantifier alternation. Taking together the above discussion of robust identifiability proofs and certifiable subgaussianity, our approach to estimate the low-degree moments of a certifiable subgaussian distribution is the following: given an ε-corrupted sample $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ from D, we want to find a (multi-)set of vectors $X' = \{x'_1, \ldots, x'_n\} \subseteq \mathbb{R}^d$ such that the uniform distributions over X and X' are ε-close in statistical distance and the uniform distribution over X' is certifiably subgaussian.

[2]We emphasize that our notion of certifiable subgaussianity is a property only of the low-degree moments of a distribution and, in this way, significantly less restrictive then the usual notion of subgaussianity (which restricts all moments of a distribution).

[3] In the special case of proving that $p(u) \geqslant 0$ holds for every vector $u \in \mathbb{S}^{d-1}$, a degree-ℓ sum-of-squares proof consists of a polynomial $q(u)$ with $\deg q \leqslant \ell - 2$ and polynomials $r_1(u), \ldots, r_t(u)$ with $\deg r_\tau \leqslant \ell/2$ such that $p = q \cdot (\|u\|^2 - 1) + r_1^2 + \cdots + r_t^2$.

It is straightforward to formulate these conditions as a system $\mathcal{E}$ of polynomial equations over the reals such that $x'_1, \ldots, x'_n$ are (some of the) variables.Our outlier-robust moment estimation proceeds by solving a standard sum-of-squares relaxation of this system $\mathcal{E}$. The solution to this relaxation can be viewed as a *pseudo-distribution* D' over vectors $x'_1, \ldots, x'_n$ that satisfies the system of equations $\mathcal{E}$. (This pseudo-distribution behaves in many ways like a classical probability distribution over vectors $x'1, \ldots, x'_n$ that satisfy the equations $\mathcal{E}$.) Our moment estimation algorithm simply outputs the expected empirical moments $\tilde{\mathbb{E}}_{D'(x'_1,\ldots,x'_n)} \frac{1}{n} \sum_{i=1}^n (x'_i)^{\otimes r}$ with respect to the pseudo-distribution D'.

We remark that previous work on computationally-efficiently outlier-robust estimation also used convex optimization techniques albeit in different ways. For example, Diakonikolas et al. solve an implicitly-defined convex optimization problem using a customized separation oracle [25]. (Their optimization problem is implicit in the sense that it is defined in terms of the uncorrupted sample which we do not observe.)

An unusual feature of the aforementioned system of equations $\mathcal{E}$ is that it also includes variables for the sum-of-squares proof of the inequality Eq. (1.2) because we want to restrict the search to those sets X' such that the uniform distribution X' is certifiably subgaussian. It is interesting to note that in this way we can use sum-of-squares as an approach to solve $\exists\forall$-problems as opposed to just the usual $\exists$-problems. (The current problem is an $\exists\forall$-problem in the sense that we want to find X' such that for all vectors u the inequality Eq. (1.2) holds for the uniform distribution over X'.)

We remark that the idea of using sum-of-squares to solve problems with quantifier alternation also plays a role in control theory (where the goal is find a dynamical system together with an associated Lyapunov functions, which can be viewed as sum-of-squares proof of the fact that the dynamical system behaves nicely in an appropriate sense). However, to the best of our knowledge, this work is the first that uses this idea for the design of computationally-efficient algorithms with provable guarantees. We remark that in a concurrent and independent work, Hopkins and Li use similar ideas to learn mixtures of well-separated spherical Gaussians [34].

2 RESULTS

2.1 Certifiably Subgaussian Distributions

In our first main result, we show that a large class of non-product distributions are also certifiably subgaussian.

Poincaré Distributions. A distribution p on $\mathbb{R}^d$ is said to be σ-*Poincaré* if for all differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ we have

$$\mathop{\mathbb{V}}_{x\sim p}[f(x)] \leqslant \sigma^2 \mathbb{E}_{x\sim p}[\|\nabla f(x)\|_2^2]. \quad (2.1)$$

This is a type of isoperimetric inequality on the distribution x and implies concentration of measure. In the full version(s), we discuss in more detail various examples of distributions that satisfy (2.1), as well as properties of such distributions. Poincaré inequalities and

distributions are intensely studied in probability theory; indeed, we rely on one such powerful result of Adamczak and Wolff Adamczak and Wolff [2] for establishing a sharp bound on the sum-of-squares algorithm's estimate of the injective norm of an i.i.d. sample from a Poincaré distribution.

Theorem 2.1. *Let p be a zero-mean σ-Poincaré distribution over $\mathbb{R}^d$ with $2t$ moment tensor $M_{2t} = \mathbb{E}_{x\sim p} x^{\otimes 2t}$. Then, for some constant $C = C(t)$ (depending only on t) with probability at least $1-\delta$ we have $\langle M_{2t}, v^{\otimes 2t}\rangle \leqslant C^{2t}\sigma\|v\|_2^{2t}$. Further, this inequality has a degree $2t$ sum of squares proof. As a consequence, isotropic σ-Poincaré distributions are C-certifiabiliy t-subgaussian for any t.*

Next, we describe our results on outlier-robust parameter estimation and clustering for certifiably subgaussian distributions.

2.2 Outlier-Robust Estimation

Without any assumptions about the underlying distribution, the best known efficient algorithms for robust mean estimation incur an estimation error that depends on the spectral norm of the covariance matrix Σ of the underlying distribution and is proportional to $\sqrt{\varepsilon}$ (where $\varepsilon > 0$ is the fraction of outliers) [26, 60]. Concretely, given an ε-corrupted sample of sufficiently larger polynomial size from a distribution D, they compute an estimate $\hat{\mu}$ for the mean μ of D such that with high probability $\|\hat{\mu}-\mu\| \leqslant O(\sqrt{\varepsilon})\cdot\|\Sigma\|^{1/2}$. Furthermore, this bound is optimal for general distributions in the sense that up to constant factors no better bound is possible information-theoretically in terms of ε and the spectral norm of Σ.

In the following theorem, we show that better bounds for the mean estimation error are possible for large classes of distributions. Concretely, we assume (certifiable) bounds on higher-order moments of the distribution (degree 4 and higher). These higher-order moment assumptions allow us to improve the estimation error as a function of ε (instead of a $\sqrt{\varepsilon}$ bound as for the unconditional mean estimation before we obtain an $\varepsilon^{1-1/k}$ if we assume a bound on the degree-k moments). Furthermore, we also obtain multiplicative approximations for the covariance matrix (in the Löwner order sense) regardless of the spectral norm of the covariance. (Note that our notion of certifiable subgaussianity does not restrict the covariance matrix in any way.)

Theorem 2.2 (Robust mean and covariance estimation under certifiable subgaussianity). *For every $C > 0$ and even $k \in \mathbb{N}$, there exists a polynomial-time algorithm that given a (corrupted) sample $S \subseteq \mathbb{R}^d$ outputs a mean-estimate $\hat{\mu} \in \mathbb{R}^d$ and a covariance-estimate $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ with the following guarantee: there exists $n_0 \leqslant (C+d)^{O(k)}$ such that if S is an ε-corrupted sample with size $|S| \geqslant n_0$ of a k-certifiably C-subgaussian distribution D over $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$, then with high probability*

$$\|\mu-\hat{\mu}\| \leqslant O(Ck)^{1/2}\cdot\varepsilon^{1-1/k}\cdot\|\Sigma\|^{1/2} \tag{2.2}$$

$$\|\Sigma^{-1/2}(\mu-\hat{\mu})\| \leqslant O(Ck)^{1/2}\cdot\varepsilon^{1-1/k} \tag{2.3}$$

$$(1-\delta)\Sigma \preceq \hat{\Sigma} \preceq (1+\delta)\Sigma \quad \text{for } \delta \leqslant O(Ck)\cdot\varepsilon^{1-2/k}. \tag{2.4}$$

For the last two bounds, we assume in addition $Ck\cdot\varepsilon^{1-2/k} \leqslant \Omega(1)$.[4]

Note that the second guarantee for the mean estimation error $\mu-\hat{\mu}$ is stronger because $\|\mu-\hat{\mu}\| \leqslant \|\Sigma\|^{1/2}\cdot\|\Sigma^{-1/2}(\mu-\hat{\mu})\|$. We remark that $\|\Sigma^{-1/2}(\mu-\hat{\mu})\|$ is the Mahalanobis distance between $\hat{\mu}$ and D.

In general, our results provide information theoretically tight way to utilize higher moment information and give optimal dependence of the error on the fraction of outliers in the input sample for every (k,ℓ)-certifiably $O(1)$-subgaussian distribution. This, in particular, improves on the mean estimation algorithm of [47] by improving on their error bound of $O(\varepsilon^{1/2})\|\Sigma\|^{1/2}\sqrt{\log(d)}$ to $O(\varepsilon^{3/4}\|\Sigma\|^{1/2}$ under their assumption of bounded 4th moments (whenever "certified" by SoS).

Frobenius vs spectral norm for covariance-matrix estimation. Previous work for robust covariance estimation [27, 47] work with Frobenius norms for measuring the estimation error $\Sigma-\hat{\Sigma}$ and obtain in this way bounds that can be stronger than ours. However, it turns out that assuming only k-certifiable subgaussianity makes it information-theoretically impossible to obtain dimension-free bounds in Frobenius norm and that we have to work with spectral norms instead. In this sense, the assumptions we make about distributions are substantially weaker compared to previous works.

Concretely, [47] show a bound of $\|\Sigma-\hat{\Sigma}\|_F \leqslant \tilde{O}(\varepsilon^{1/2})\|\Sigma\|$ assuming a 4-th moment bound and that the distribution is an affine transformation of a product distribution. [27] show a bound $\|\Sigma-\hat{\Sigma}\|_F \leqslant O(\varepsilon)\|\Sigma\|$ assuming the distribution is Gaussian.[5]

However, even for simple k-certifiably subgaussian distributions with parameter $C \leqslant O(1)$, the information-theoretically optimal error in terms of Frobenius norm is $\|\Sigma-\hat{\Sigma}\|_F \leqslant O(\sqrt{d}\cdot\varepsilon^{1-2/k})\cdot\|\Sigma\|$. Concretely, consider the mixture of $\mathcal{N}(0,I)$ and $\mathcal{N}(0,\varepsilon^{-2/k}I)$ with weights $1-\varepsilon$ and ε, respectively. Then, it's easy to confirm that both the mixture and the standard gaussian $\mathcal{N}(0,I)$ are k-certifiably 2-subgaussian and at most ε far in total variation distance. Thus, given only k-certifiable 2-subgaussianity, it is information theoretically impossible to decide which of the above two distributions generated a given ε-corrupted sample. Finally, the difference of their covariance equals $\varepsilon(\varepsilon^{-2/k}-1)I$ which has Frobenius norm $\Omega(\varepsilon^{1-2/k})\sqrt{d}$ for any $\varepsilon < 1$. For this case, our algorithms guarantee the significantly stronger[6] spectral-norm bound $\|\Sigma-\hat{\Sigma}\| \leqslant O(\varepsilon^{1-2/k})\cdot\|\Sigma\|$.

Multiplicative vs. additive estimation error. Another benefit of our covariance estimation algorithm is that provide a multiplicative

[4] This notation means that we require $Ck\cdot\varepsilon^{1-2/k} \leqslant c_0$ for some absolute constant $c_0 > 0$ (that could in principle be extracted from the proof).

[5] In the Gaussian case, [27] establish the stronger bound $\|\Sigma\|^{-1/2}(\Sigma-\hat{\Sigma})\Sigma^{-1/2}{}_F \leqslant O(\varepsilon)$. This norm can be viewed as the Frobenius norm in a transformed space. Our bounds are for the spectral norm in the same transformed space.

[6] Similarly to [27], we also work with norms in a transformed space (in which the distribution is isotropic) and obtain the stronger bound $\|\Sigma^{-1/2}(\Sigma-\hat{\Sigma})\Sigma^{-1/2}\| \leqslant O(\varepsilon^{1-2/k})$.

approximation guarantee, that is, the quadratic form of the estimated covariance at any vector u is within $(1 \pm \delta)$ of the quadratic form of the true covariance. This strong guarantee comes in handy, for example, in *whitening* or computing an isotropic transformation of the data—a widely used primitive in algorithm design. Indeed, this ability to use the estimated covariance to whiten the data is crucial in our outlier-robust algorithm for independent component analysis.The Frobenius norm error guarantees, in general, do not imply good multiplicative approximations and thus cannot be used for this application.

We note that in the special case of gaussian distributions, the results of Diakonikolas et al. [25] allow recovering the mean and covariance in fixed polynomial time with better dependence of the error on the fraction of outliers that grows as $\tilde{O}(\varepsilon)\|\Sigma\|^{1/2}$. Our results when applied for the special case of gaussian mean and covariance estimation will require a running time of $d^{O(\sqrt{\log(1/\varepsilon)})}$ to achieve a similar error guarantee. Their algorithm for gaussian covariance estimation also provides multiplicative error guarantees.

Robust Estimation of Higher Moments. Our techniques for robustly estimating mean and covariance of a distribution extend in a direct way to robustly estimating higher-order moment tensors. In order to measure the estimation error, we use a variant of the injective tensor norm (with respect to a transformed space where the distribution is isotropic), which generalizes the norm we use for the estimation error of the covariance matrix. This error bound allows us to estimate for every direction $u \in \mathbb{R}^d$, the low-degree moments of the distribution in direction u with small error compared to the second moment in direction u.

The approaches in previous works fact inherent obstacles in generalizing to the problem of estimating the higher moments with multiplicative (i.e. in every direction u) error guarantees. This type of error is in fact crucial in applications for learning latent variable models such as mixtures of Gaussians and independent component analysis.

In fact, our guarantees are in some technical way stronger, which is crucial for our applications of higher-order moment estimates. Unlike spectral norms, injective norms are NP-hard to compute (even approximately, under standard complexity assumptions). For this reason, it is not clear how to make use of an injective-norm guarantee when processing moment-estimates further. Fortunately, it turns out that our algorithm not only guarantees an injective-norm bound for the error but also a good certificate for this bound, in form of a low-degree sum-of-squares proof. It turns out that this kind of certificate is precisely what we need for our applications—in particular, recent tensor decomposition algorithms based on sum-of-squares [49] can tolerate errors with small injective norm if that is certified by a low-degree sum-of-squares proof.

Theorem 2.3 (Robust Higher Moment Estimation). *For every $C > 0$ and even $k \in \mathbb{N}$, there exists a polynomial-time algorithm that given a (corrupted) sample $S \subseteq \mathbb{R}^d$ outputs a moment-estimates $\hat{M}_2 \in \mathbb{R}^{d^2}, \ldots, \hat{M}_k \in \mathbb{R}^{d^k}$ with the following guarantee: there exists $n_0 \leqslant (C+d)^{O(k)}$ such that if S is an ε-corrupted sample with size $|S| \geqslant n_0$ of a k-certifiably C-subgaussian distribution D over $\mathbb{R}^d$ with moment-tensors $M_2 \in \mathbb{R}^{d^2}, \ldots, M_k \in \mathbb{R}^{d^k}$ such that $Ck \cdot \varepsilon^{1-2/k} \leqslant \Omega(1)$, then with high probability for every $r \leqslant k/2$,*

$$\forall u \in \mathbb{R}^d . \langle M_r - \hat{M}_r, u^{\otimes r}\rangle^2 \leqslant \delta_r \cdot \langle M_2, u^{\otimes 2}\rangle^r \quad \text{for } \delta_r \leqslant O(Ck)^{r/2} \cdot \varepsilon^{1-\frac{r}{k}} \tag{2.5}$$

Furthermore, there exist degree-k sum-of-squares proofs of the above polynomial inequalities in u.

Information-theoretic optimality. We show that the error guarantees in our robust moment-estimation algorithms are tight in their dependence on both k and ε. For example, we show that there are two k-certifiably $O(1)$-subgaussian distributions with statistical distance ε but means that are $\Omega(\sqrt{k}\varepsilon^{1-1/k})$ apart. A similar statement holds for higher-order moments. The distributions are just mixtures of two one-dimensional Gaussians.

Application: independent component analysis. As an immediate application of our robust moment estimation algorithm, we get an algorithm for Outlier Robust Independent Component Analysis. Independent component analysis (also known as blind source separation) is a fundamental problem in signal processing, machine learning and theoretical computer science with applications to diverse areas including neuroscience. Lathauwer et. al. [22], following up on a long line of work gave algorithms for ICA based on 4th order tensor decomposition. A noise-tolerant version of this algorithm was developed in [49]. There is also a line of work in theoretical computer science on designing efficient algorithms for ICA [30, 64].

In the ICA problem, we are given a non-singular[7] mixing matrix $A \in \mathbb{R}^{d\times d}$ with condition number κ and a product distribution on $\mathbb{R}^d$. The observations come from the model $\{Ax\}$ that is, the observed samples are linear transformations (using the mixing matrix) of independent draws of the product random variable x. The goal is to recover columns of A up to small relative error in Euclidean norm (up to signs and permutations) from samples. It turns out that information theoretic recovery of A is possible whenever at most one source is non-gaussian. A widely used convention in this regard is the 4th moment assumption: for each i, $\mathbb{E}[x_i^4] \neq 3\,\mathbb{E}[x_i^2]$. It turns out that we can assume $\mathbb{E}[x_i^2] = 1$ without loss of generality so this condition reduces to asserting $\mathbb{E}[x_i^4] \neq 3$.

Outlier robust version of ICA was considered as an application of the outlier-robust mean and covariance estimation problems in [47]. They sketched an algorithm with the guarantee that the relative error in the columns of A is at most $\varepsilon\sqrt{\log(d)}\,\mathrm{poly}(\kappa)$.

In particular, this guarantee is meaningful only if the fraction of outliers $\varepsilon \ll \frac{1}{\sqrt{\log(d)}\,\mathrm{poly}(\kappa)}$. Here, we improve upon their result by giving an outlier-robust ICA algorithm that recovers columns

[7] We could also consider the case that A is rectangular and its columns are linearly independent. Essentially the same algorithm and analysis would go through in this case. We focus on the quadratic case for notational simplicity.

of A up to an error that is *independent* of both dimension d and the condition number κ of the mixing matrix A.

Our algorithm directly follows by applying 4th order tensor decomposition. However, a crucial step in the algorithm involves "whitening" the 4th moments by using an estimate of the covariance matrix. Here, the multiplicative guarantees obtained in estimating the covariance matrix are crucial - estimates with respect to the Frobenius norm error do not give such whitening transformation in general. This whitening step essentially allows us to pretend that the mixing matrix A is well-conditioned leading to no dependence on the condition number in the error.

Theorem 2.4 (Robust independent component analysis). *For every $C \geqslant 1$ and even $k \in \mathbb{N}$, there exists a polynomial-time algorithm that given a (corrupted) sample $S \subseteq \mathbb{R}^d$ outputs component estimates $\hat{a}_1, \ldots, \hat{a}_d \in \mathbb{R}^d$ with the following guarantees: Suppose $A \in \mathbb{R}^{d\times d}$ is a non-singular matrix with condition number κ and columns $a_1, \ldots, a_d \in \mathbb{R}^d$. Suppose $\mathbf{x}$ is a centered random vector with d independent coordinates such that every coordinate $i \in [d]$ satisfies $\mathbb{E}[\mathbf{x}_i^2] = 1$, $\mathbb{E}[\mathbf{x}_i^4] - 3 = \gamma \neq 0$, and $\mathbb{E}[\mathbf{x}_i^k]^{1/k} \leqslant \sqrt{Ck}$. Then, if S is an ε-corrupted sample of size $|S| \geqslant n_0$ from the distribution $\{A\mathbf{x}\}$, where $n_0 \leqslant (C + \kappa + d)^{O(k)}$, the component estimates satisfy with high probability*

$$\max_{\pi\in S_d} \min_{i\in[d]} \langle A^{-1}\hat{a}_i, A^{-1}a_{\pi(i)}\rangle^2 \geqslant 1-\delta \quad \text{for } \delta < (1+\tfrac{1}{|\gamma|})\cdot O(C^2k^2)\cdot\varepsilon^{1-4/k}. \tag{2.6}$$

The quantity $\langle A^{-1}\hat{a}_i, A^{-1}a_{\pi(i)}\rangle$ is closely related to the Mahalanobis distance between $\hat{a}_i$ and $a_{\pi(i)}$ with respect to the distribution $\{A\mathbf{x}\}$

Application: learning mixtures of Gaussians. As yet another immediate application of our robust moment estimation algorithm, we get an outlier-robust algorithm for learning mixtures of spherical Gaussians. Our algorithm works under the assumption that the means are linearly independent (and that the size of the sample grows with their condition number). In return, our algorithm does not require the means of the Gaussians to be well-separated. Our algorithm can be viewed as an outlier-robust version of tensor-decomposition based algorithms for mixtures of Gaussians [15, 37].

Theorem 2.5 (Robust estimation of mixtures of spherical Gaussians). *Let D be mixtures of $\mathcal{N}(\mu_i, I)$ for $i \leqslant q$ with uniform[8] mixture weights. Assume that μ_is are linearly independent and, further, assume that κ, the smallest non-zero eigenvalue of $\frac{1}{q}\sum_i \mu_i\mu_i^\top$ is $\Omega(1)$.*

Given an ε-corrupted sample of size $n \geqslant n_0 = \Omega((d\log(d))^{k/2}/\varepsilon^2)$, for every $k \geqslant 4$, there's a $\mathrm{poly}(n)d^{O(k)}$ time algorithm that recovers $\hat{\mu}_1, \hat{\mu}_2, \ldots \hat{\mu}_q$ so that there's a permutation $\pi : [q] \to [q]$ satisfying

$$\max_i \|(\frac{1}{q}\sum_i \mu_i\mu_i^\top)^{-1/2}(\hat{\mu}_i - \mu_{\pi(i)})\| \leqslant O(qk)\varepsilon^{1/3-1/k}.$$

[8]While our algorithm generalizes naturally to arbitrary mixture weights, we restrict to this situation for simplicity

Diakonikolas et. al. [25] gave an outlier-robust algorithm that learns mixtures of q gaussians with error $\approx q\sqrt{\varepsilon}$ in each of the recovered means. Their algorithm is polynomial in the dimension but has an exponential dependence on number of components q in the running time. Under the additional assumption that the means are linearly independent, our algorithm (say for $k = 8$) recovers similar error guarantees as theirs but runs in time polynomial in both q and d. The key difference is the power of our algorithm to recover a multiplicative approximation to the 4th moment tensor which allows us to apply blackbox tensor decomposition based methods and run in fixed polynomial time [39].

2.3 Distribution-Agnostic Robust Clustering

Specifically, we show that for any $\gamma > 0$, given a balanced mixture of k Poincaré distributions with means separated by $\Omega(k^\gamma)$, we can successfully cluster n samples from this mixture in $n^{O(1/\gamma)}$ time (by using $O(1/\gamma)$ levels of the sum-of-squares hierarchy). Similarly, given samples from a Poincaré distribution with an ε fraction of adversarial corruptions, we can estimate its mean up to an error of $O(\varepsilon^{1-\gamma})$ in $n^{O(1/\gamma)}$ time. In fact, we will see below that we get both at once: a robust clustering algorithm that can learn well-separated mixtures even in the presence of arbitrary outliers.

To our knowledge such a result was not previously known even in the second-moment case (Charikar et al. [20] and Steinhardt et al. [61] study this setting but only obtain results in the *list-decodable* learning model). Our result only relies on the SOS-certifiability of the moment tensor, and holds for any deterministic point set for which such a sum-of-squares certificate exists.

Despite their generality, our results are strong enough to yield new bounds even in very specific settings such as learning balanced mixtures of k spherical Gaussians with separation $\Omega(k^\gamma)$. Our algorithm allows recovering the true means in $n^{O(1/\gamma)}$ time and partially resolves an open problem posed in the recent work of Regev and Vijayaraghavan [59].

Certifying injective norms of moment tensors appears to be a useful primitive and could help enable further applications of the sum of squares method in machine learning. Indeed, [43] studies the problem of robust estimation of higher moments of distributions that satisfy a bounded-moment condition closely related to approximating injective norms. Their relaxation and the analysis are significantly different from the present work; nevertheless, our result for Poincaré distributions immediately implies that the robust moment estimation algorithm of [43] succeeds for a large class of Poincaré distributions.

Our first main result regards efficient upper bounds on the injective norm of the moment tensor of any Poincaré distribution. Let $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ be n i.i.d. samples from a Poincaré distribution with mean μ, and let $M_{2t} = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^{\otimes 2t}$ be the empirical estimate of the $2t$th moment tensor. We are interested in upper-bounding the injective norm, which can be equivalently expressed

in terms of the moment tensor as

$$\sup_{\|v\|_2\leqslant 1} \frac{1}{n}\sum_{i=1}^{n}\langle x_i-\mu, v\rangle^{2t} = \sup_{\|v\|_2\leqslant 1} \langle M_{2t}, v^{\otimes 2t}\rangle. \qquad (2.7)$$

Standard results yield dimension-free upper bounds on (2.7) for all Poincaré distributions. Our first result is a "sum-of-squares proof" of this fact giving an *efficient* method to certify dimension-free upper bounds on (2.7) for samples from any Poincaré distribution.

Specifically, let the sum of squares norm of M_{2t}, denoted by $\|M_{2t}\|_{\mathrm{sos}_{2t}}$, be the degree-$2t$ sum-of-squares relaxation of (2.7) (we discuss such norms and the sum-of-squares method in more detail in the full version) for now the important fact is that $\|M_{2t}\|_{\mathrm{sos}_{2t}}$ can be computed in time $(nd)^{O(t)}$). We show that for a large enough sample from a distribution that satisfies the Poincaré inequality, the sum-of-squares norm of the moment tensor is upper bounded by a dimension-free constant.

Theorem 2.6. *Let p be a σ-Poincaré distribution over $\mathbb{R}^d$ with mean μ. Let $x_1,\ldots,x_n \sim p$ with $n \geqslant (2d\log(dt/\delta))^t$. Then, for some constant C_t (depending only on t) with probability at least $1-\delta$ we have $\|M_{2t}\|_{\mathrm{sos}_{2t}} \leqslant C_t\sigma$, where $M_{2t} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^{\otimes 2t}$.*

As noted above, previous sum-of-squares bounds worked for specialized cases such as product distributions. Theorem 2.6 is key to our applications that crucially rely on 1) going beyond product distributions and 2) using sos_{2t} norms as a proxy for injective norms for higher moment tensors.

Outlier-Robust Agnostic Clustering. Our second main result is an efficient algorithm for *outlier-robust agnostic clustering* whenever the "ground-truth" clusters have moment tensors with bounded sum-of-squares norms.

Concretely, the input is data points $x_1,\ldots,x_n$ of n points in $\mathbb{R}^d$, a $(1-\varepsilon)$ fraction of which admit a (unknown) partition into sets $I_1,\ldots,I_k$ each having bounded sum-of-squares norm around their corresponding means $\mu_1,\ldots,\mu_k$. The remaining ε fraction can be arbitrary outliers. Observe that in this setting, we do not make any explicit distributional assumptions.

We will be able to obtain strong estimation guarantees in this setting so long as the clusters are *well-separated* and the fraction ε of outliers is not more than $\alpha/8$, where α is the fraction of points in the smallest cluster. We define the separation as $\Delta = \min_{i\neq j}\|\mu_i-\mu_j\|_2$. A lower bound on Δ is information theoretically necessary even in the special case of learning mixtures of identity-covariance gaussians without any outliers (see [59]).

Theorem 2.7. *Suppose points $x_1,\ldots,x_n \in \mathbb{R}^d$ can be partitioned into sets $I_1,\ldots,I_k$ and* out*, where the I_j are the clusters and* out *is a set of outliers of size εn. Suppose I_j has size $\alpha_j n$ and mean μ_j, and that its $2t$th moment $M_{2t}(I_j)$ satisfies $\|M_{2t}(I_j)\|_{\mathrm{sos}_{2t}} \leqslant B$. Also suppose that $\varepsilon \leqslant \alpha/8$ for $\alpha = \min_{j=1}^{k}\alpha_j$.*

Finally, suppose the separation $\Delta \geqslant C_{sep}\cdot B/\alpha^{1/t}$, with $C_{sep} \geqslant C_0$ (for a universal constant C_0). Then there is an algorithm running in time $(nd)^{O(t)}$ and outputting means $\hat{\mu}_1,\ldots,\hat{\mu}_k$ such that $\|\hat{\mu}_j-\mu_j\|_2 \leqslant O(B(\varepsilon/\alpha + C_{sep}^{-2t})^{1-1/2t})$ for all j.

The parameter B specifies a bound on the variation in each cluster. The separation condition says that the distance between cluster means must be slightly larger (by a $\alpha^{-1/t}$ factor) than this variation. The error in recovering the cluster means depends on two terms—the fraction of outliers ε, and the separation C_{sep}.

To understand the guarantees of the theorem, let's start with the case where $\varepsilon = 0$ (no outliers) and $\alpha = 1/k$ (all clusters have the same size). In this case, the separation requirement between the clusters is $B\cdot k^{1/t}$ where B is the bound on the moment tensor of order $2t$. The theorem guarantees a recovery of the means up to an error in Euclidean norm of $O(B)$. By taking t larger (and spending the correspondingly larger running time), our clustering algorithm works with separation k^γ for any constant γ. This is the first result that goes beyond the separation requirement of $k^{1/2}$ in the *agnostic* clustering setting—i.e., without making distributional assumptions on the clusters.

It is important to note that even in 1 dimension, it is information theoretically impossible to recover cluster means to an error $\ll B$ when relying only on $2t$th moment bounds. A simple example to illustrate this is obtained by taking a mixture of two distributions on the real line with bounded $2t$th moments but small overlap in the tails. In this case, it is impossible to correctly classify the points that come from the the overlapping part. Thus, a fraction of points in the tail always end up misclassified, shifting the true means. The recovery error of our algorithm does indeed drop as the separation (controlled by C_{sep}) between the true means increases (making the overlapping parts of the tail smaller). We note that for the specific case of spherical gaussians, we can exploit their parametric structure to get arbitrarily accurate estimates even for fixed separation; see Corollary 2.9.

Next, let's consider $\varepsilon \neq 0$. In this case, if $\varepsilon \ll \alpha$, we recover the means up to an error of $O(B)$ again (for $C_{sep} \geqslant C_0$). It is intuitive that the recovery error for the means should grow with the number of outliers, and the condition $\varepsilon \leqslant \alpha/8$ is necessary, as if $\varepsilon \geqslant \alpha$ then the outliers could form an entirely new cluster making recovery of the means information-theoretically impossible.

We also note that in the degenerate case where $k = 1$ (a single cluster), Theorem 2.7 yields results for robust mean estimation of a set of points corrupted by an ε fraction of outliers. In this case we are able to estimate the mean to error $\varepsilon^{\frac{2t-1}{2t}}$; when $t = 1$ this is $\sqrt{\varepsilon}$, which matches the error obtained by methods based on second moments [24, 46, 61]. For $t = 2$ we get error $\varepsilon^{3/4}$, for $t = 3$ we get error $\varepsilon^{5/6}$, and so on, approaching an error of ε as $t \to \infty$. In particular, this pleasingly approaches the rate $\tilde{O}(\varepsilon)$ obtained by much more bespoke methods that rely strongly on specific distributional assumptions [23, 46].

Note that we could not hope to do better than $\varepsilon^{\frac{2t-1}{2t}}$, as that is the information-theoretically optimal error for distributions with

bounded $2t$th moments (even in one dimension), and degree-$2t$ SOS only "knows about" moments up to $2t$.

Finally, we can obtain results even for clusters that are not well-separated, and for fractions of outliers that could exceed α. In this case we no longer output exactly k means, and must instead consider the list-decodable model Balcan et al. [8], Charikar et al. [20], where we output a list of $O(1/\alpha)$ means of which the true means are a sublist.

Applications. Putting together Theorem 2.6 and Theorem 2.7 immediately yields corollaries for learning mixtures of Poincaré distributions, and in particular mixtures of Gaussians.

Corollary 2.8 (Disentangling Mixtures of Arbitrary Poincaré Distributions). *Suppose that we are given a dataset of n points $x_1, \dots, x_n$, such that at least $(1-\varepsilon)n$ points are drawn from a mixture $\alpha_1 p_1 + \cdots + \alpha_k p_k$ of k distributions, where p_j is σ-Poincaré with mean μ_j (the remaining εn points may be arbitrary). Let $\alpha = \min_{j=1}^k \alpha_j$. Also suppose that the separation Δ is at least $C_{sep} \cdot C_t \sigma / \alpha^{1/t}$, for some constant C_t depending only on t and some $C_{sep} \geqslant 1$.*

Then, assuming that $\varepsilon \leqslant \frac{\alpha}{10}$, for some $n = O((2d\log(tkd/\delta))^t/\alpha + d\log(k/\delta)/\alpha\varepsilon^2)$, there is an algorithm running in $n^{O(t)}$ time which with probability $1-\delta$ outputs candidate means $\hat{\mu}_1, \dots, \hat{\mu}_k$ such that $\|\hat{\mu}_j - \mu_j\|_2 \leqslant C_t' \sigma (\varepsilon/\alpha + C_{sep}^{-2t})^{\frac{2t-1}{2t}}$ for all j (where C_t' is a different universal constant).

The $1/\alpha$ factor in the sample complexity is so that we have enough samples from every single cluster for Theorem 2.6 to hold. The extra term of $d\log(k/\delta)/\varepsilon^2$ in the sample complexity is so that the empirical means of each cluster concentrate to the true means.

Corollary 2.8 is one of the strongest results on learning mixtures that one could hope for. If the mixture weights α are all at least $1/\operatorname{poly}(k)$, then Corollary 2.8 implies that we can cluster the points as long as the separation $\Delta = \Omega(k^\gamma)$ for any $\gamma > 0$. Even for spherical Gaussians the best previously known algorithms required separation $\Omega(k^{1/4})$. On the other hand, Corollary 2.8 applies to a large family of distributions including arbitrary strongly log-concave distributions. Moreover, while the Poincaré inequality does not directly hold for discrete distributions, a large class of discrete distributions, including product distributions over bounded domains, will satisfy the Poincaré inequality after adding zero-mean Gaussian noise.

For mixtures of Gaussians in particular, we can do better, and in fact achieve vanishing error independent of the separation:

Corollary 2.9 (Learning Mixtures of Gaussians). *Suppose that $x_1, \dots, x_n \in \mathbb{R}^d$ are drawn from a mixture of k Gaussians: $p = \sum_{j=1}^k \alpha_j \mathcal{N}(\mu_j, I)$, where $\alpha_j \geqslant 1/\operatorname{poly}(k)$ for all j. Then for any $\gamma > 0$, there is a separation $\Delta_0 = O(k^\gamma)$ such that given $n \geqslant \operatorname{poly}(d^{1/\gamma}, k, 1/\varepsilon)\log(k/\delta)$ samples from p, if the separation $\Delta \geqslant \Delta_0$, then with probability $1-\delta$ we obtain estimates $\hat{\mu}_1, \dots, \hat{\mu}_k$ with $\|\hat{\mu}_j - \mu_j\|_2 \leqslant \varepsilon$ for all j.*

Remark 2.10. This partially resolves an open question of Regev and Vijayaraghavan [59], who ask whether it is possible to efficiently learn mixtures of Gaussians with separation $\sqrt{\log k}$.

The error now goes to 0 as $n \to \infty$, which is not true in the more general Corollary 2.8. This requires invoking Theorem IV.1 of Regev and Vijayaraghavan [59], which, given a sufficiently good initial estimate of the means of a mixture of Gaussians, shows how to get an arbitrarily accurate estimate. As discussed before, such a result is specific to Gaussians and in particular is information-theoretically impossible for mixtures of general Poincaré distributions.

Proof Sketch and Technical Contributions. We next sketch the proofs of our two main theorems (Theorem 2.6 and Theorem 2.7) while indicating which parts involve new technical ideas.

2.3.1 Sketch of Theorem 2.6. For simplicity, we will only focus on SOS-certifiability in the infinite-data limit, i.e. on showing that SOS can certify an upper bound $\mathbb{E}_{x\sim p}[\langle x-\mu, v\rangle^{2t}] \leqslant C_t \sigma^{2t} \|v\|_2^{2t}$. (In full version, we will show that finite-sample concentration follows due to the matrix Rosenthal inequality [50].)

We make extensive use of a result of Adamczak and Wolff [2]; it is a very general result on bounding non-Lipschitz functions of Poincaré distributions, but in our context the important consequence is the following:

> If $f(x)$ is a degree-t polynomial such that $\mathbb{E}_p[\nabla^j f(x)] = 0$ for $j = 0, \dots, t-1$, then $\mathbb{E}_p[f(x)^2] \leqslant C_t \sigma^{2t} \|\nabla^t f(x)\|_F^2$ for a constant C_t, assuming p is σ-Poincaré. (Note that $\nabla^t f(x)$ is a constant since f is degree-t.)

Here $\|A\|_F^2$ denotes the Frobenius norm of the tensor A, i.e. the ℓ_2-norm of A if it were flattened into a d^t-element vector.

We can already see why this sort of bound might be useful for $t = 1$. Then if we let $f_v(x) = \langle x-\mu, v\rangle$, we have $\mathbb{E}[f_v(x)] = 0$ and hence $\mathbb{E}_p[\langle x-\mu, v\rangle^2] \leqslant C_1\sigma^2\|v\|_2^2$. This exactly says that p has bounded covariance.

More interesting is the case $t = 2$. Here we will let $f_A(x) = \langle (x-\mu)(x-\mu)^\top - \Sigma, A\rangle$, where μ is the mean and Σ is the covariance of p. It is easy to see that both $\mathbb{E}[f_A(x)] = 0$ and $\mathbb{E}[\nabla f_A(x)] = 0$. Therefore, we have $\mathbb{E}[\langle (x-\mu)(x-\mu)^\top - \Sigma, A\rangle^2] \leqslant C_2\sigma^4\|A\|_F^2$.

Why is this bound useful? It says that if we unroll $(x-\mu)(x-\mu)^\top - \Sigma$ to a d^2-dimensional vector, then this vector has bounded covariance (since if we project along any direction A with $\|A\|_F = 1$, the variance is at most $C_2\sigma^4$). This is useful because it turns out sum-of-squares "knows about" such covariance bounds; indeed, this type of covariance bound is exactly the property used in Barak et al. [9] to certify 4th moment tensors over the hypercube. In our case it yields a sum-of-squares proof that $\mathbb{E}[\langle (x-\mu)^{\otimes 4} - \Sigma^{\otimes 2}, v^{\otimes 4}\rangle] \preceq_{sos} C_2\sigma^4\|v\|_2^4$, which can then be used to bound the 4th moment $\mathbb{E}[\langle x-\mu, v\rangle^4] = \mathbb{E}[\langle (x-\mu)^{\otimes 4}, v^{\otimes 4}\rangle]$.

Motivated by this, it is natural to try the same idea of "subtracting off the mean and squaring" with $t = 4$. Perhaps we could define $f_A(x) = \langle ((x-\mu)^{\otimes 2} - \Sigma)^{\otimes 2} - \mathbb{E}[((x-\mu)^{\otimes 2} - \Sigma)^{\otimes 2}], A\rangle$?

Alas, this does not work—while there is a suitable polynomial $f_A(x)$ for $t = 4$ that yields sum-of-squares bounds, it is somewhat more subtle. For simplicity we will write the polynomial for $t = 3$. It is the following: $f_A(x) = \langle (x-\mu)^{\otimes 3} - 3(x-\mu)\otimes\Sigma - M_3, A\rangle$, where $M_3 = \mathbb{E}[(x-\mu)^{\otimes 3}]$ is the third-moment tensor of p. By checking that $\mathbb{E}[f_A(x)] = \mathbb{E}[\nabla f_A(x)] = \mathbb{E}[\nabla^2 f_A(x)] = 0$, we obtain that the tensor $F_3(x) = (x-\mu)^{\otimes 3} - 3(x-\mu)\otimes\Sigma - M_3$, when unrolled to a d^3-dimensional vector, has bounded covariance, which means that sum-of-squares knows that $\mathbb{E}[\langle F_3(x)^{\otimes 2}, v^{\otimes 6}\rangle]$ is bounded for all $\|v\|_2 \leqslant 1$.

However, this is not quite what we want—we wanted to show that $\mathbb{E}[\langle (x-\mu)^{\otimes 6}, v^{\otimes 6}\rangle]$ is bounded. Fortunately, the leading term of $F_3(x)^{\otimes 2}$ is indeed $(x-\mu)^{\otimes 6}$, and all the remaining terms are lower-order. So, we can subtract off $F_3(x)$ and recursively bound all of the lower-order terms to get a sum-of-squares bound on $\mathbb{E}[\langle (x-\mu)^{\otimes 6}, v^{\otimes 6}\rangle]$. The case of general t follows similarly, by carefully constructing a tensor $F_t(x)$ whose first $t-1$ derivatives are all zero in expectation.

There are a couple contributions here beyond what was known before. The first is identifying appropriate tensors $F_t(x)$ whose covariances are actually bounded so that sum-of-squares can make use of them. For $t = 1, 2$ (the cases that had previously been studied) the appropriate tensor is in some sense the "obvious" one $(x-\mu)^{\otimes 2} - \Sigma$, but even for $t = 3$ we end up with the fairly non-obvious tensor $(x-\mu)^{\otimes 3} - 3(x-\mu)\otimes\Sigma - M_3$. (For $t = 4$ it is $(x-\mu)^{\otimes 4} - 6(x-\mu)^{\otimes 2}\otimes\Sigma - 4(x-\mu)\otimes M_3 - M_4 + 6\Sigma\otimes\Sigma$.) While these tensors may seem mysterious a priori, they are actually the unique tensor polynomials with leading term $x^{\otimes t}$ such that all derivatives of order $j < t$ have mean zero. Even beyond Poincaré distributions, these seem like useful building blocks for sum-of-squares proofs.

The second contribution is making the connection between Poincaré distributions and the above polynomial inequalities. The well known work of Latała Latała [48] establishes non-trivial estimates of upper bounds on the moments of polynomials of Gaussians, of which the inequalities used here are a special case. Adamczak and Wolff [2] show that these inequalities also hold for Poincaré distributions. However, it is not a priori obvious that these inequalities should lead to sum-of-squares proofs, and it requires a careful invocation of the general inequalities to get the desired results in the present setting.

2.3.2 Sketch of Theorem 2.7. We next establish our result on robust clustering. In fact we will establish a robust mean estimation result which will lead to the clustering result—specifically, we will show that if a set of points $x_1, \ldots, x_n$ contains a subset $\{x_i\}_{i\in I}$ of size αn that is SOS-certifiable, then the mean (of the points in I) can be estimated regardless of the remaining points. There are two parts: if $\alpha \approx 1$ we want to show error going to 0 as $\alpha \to 1$, while if $\alpha \ll 1$ we want to show error that does not grow too fast as $\alpha \to 0$. In the latter case we will output $O(1/\alpha)$ candidates for the mean and show that at least one of them is close to the true mean (think of these candidates as accounting for $O(1/\alpha)$ possible clusters in the data). We will later prune down to exactly k means for well-separated clusters.

For $t = 1$ (which corresponds to bounded covariance), the $\alpha \to 0$ case is studied in Charikar et al. [20]. A careful analysis of the proof there reveals that all of the relevant inequalities are sum-of-squares inequalities, so there is a sum-of-squares generalization of the algorithm in Charikar et al. [20] that should give bounds for SOS-certifiable distributions. While this would likely lead to some robust clutering result, we note the bounds we achieve here are stronger than those in Charikar et al. [20], as Charikar et al. [20] do not achieve tight results when the clusters are well-separated. Moreover, the proof in Charikar et al. [20] is complex and would be somewhat tedious to extend in full to the sum-of-squares setting.

We combine and simplify ideas from both Charikar et al. [20] and Steinhardt et al. [61] to obtain a relatively clean algorithm. In fact, we will see that a certain mysterious constraint appearing in Charikar et al. [20] is actually the natural constraint from a sum-of-squares perspective.

Our algorithm is based on the following optimization. Given points $x_1, \ldots, x_n$, we will try to find points $w_1, \ldots, w_n$ such that $\frac{1}{n}\sum_{i=1}^n \tilde{\mathbb{E}}_{\xi(v)}[\langle x_i - w_i, v\rangle^{2t}]$ is small for all pseudodistributions ξ over the sphere. This is natural because we know that for the good points x_i and the true mean μ, $\langle x_i - \mu, v\rangle^{2t}$ is small (by the SOS-certifiability assumption). However, without further constraints this is not a very good idea because the trivial optimum is to set $w_i = x_i$. We would somehow like to ensure that the w_i cannot overfit too much to the x_i; it turns out that the natural way to measure this degree of overfitting is via the quantity $\sum_{i\in I}\langle w_i - \mu, w_i\rangle^{2t}$.

Of course, this quantity is not known because we do not know μ. But we *do* know that $\sum_{i\in I}\tilde{\mathbb{E}}_{\xi(v)}[\langle w_i - \mu, v\rangle^{2t}]$ is small for all pseudodistributions (because the corresponding quantity is small for $x_i - \mu$ and $w_i - x_i$, and hence also for $w_i - \mu = (w_i - x_i) + (x_i - \mu)$ by Minkowski's inequality). Therefore, we impose the following constraint: *whenever* $z_1, \ldots, z_n$ *are such that* $\sum_{i=1}^n \tilde{\mathbb{E}}_\xi[\langle z_i, v\rangle^{2t}] \leqslant 1$ *for all* ξ*, it is also the case that* $\sum_{i=1}^n \langle z_i, w_i\rangle^{2t}$ *is small.* This constraint is not efficiently imposable, but it does have a simple sum-of-squares relaxation. Namely, we require that $\sum_{i=1}^n \langle Z_i, w_i^{\otimes 2t}\rangle$ is small whenever $Z_1, \ldots, Z_n$ are pseudomoment tensors satisfying $\sum_{i=1}^n Z_i \preceq_{sos} I$.

Together, this leads to seeking $w_1, \ldots, w_n$ such that

$$\sum_{i=1}^n \tilde{\mathbb{E}}_\xi[\langle x_i - w_i, v\rangle^{2t}] \text{ is small for all } \xi, \text{ and}$$
$$\sum_{i=1}^n \langle Z_i, w_i^{\otimes 2t}\rangle \text{ is small whenever } \sum_i Z_i \preceq_{sos} I. \quad (2.8)$$

If we succeed in this, we can show that we end up with a good estimate of the mean (more specifically, the w_i are clustered into a small number of clusters, such that one of them is centered near μ). The above is a convex program, and thus, if this is impossible, by duality there must exist specific ξ and $Z_1, \ldots, Z_n$ such that the above quantities cannot be small for *any* $w_1, \ldots, w_n$. But for fixed

ξ and $Z_{1:n}$, the different w_i are independent of each other, and in particular it should be possible to make both sums small at least for the terms coming from the good set I. This gives us a way of performing *outlier removal*: look for terms where $\min_w \tilde{\mathbb{E}}_\xi[\langle x_i - w, v\rangle^{2t}]$ or $\min_w \langle Z_i, w\rangle$ is large, and remove those from the set of points. We can show that after a finite number of iterations this will have successfully removed many outliers and few good points, so that eventually we must succeed in making both sums small and thus get a successful clustering.

Up to this point the proof structure is similar to Steinhardt et al. [61]; the main innovation is the constraint involving the z_i, which bounds the degree of overfitting. In fact, when $t = 1$ this constraint is the dual form of one appearing in Charikar et al. [20], which asks that $w_i^{\otimes 2} \preceq Y$ for all i, for some matrix Y of small trace. In Charikar et al. [20], the matrix Y couples all of the variables, which complicates the analysis. In the form given here, we avoid the coupling and also see why the constraint is the natural one for controlling overfitting.

To finish the proof, it is also necessary to iteratively re-cluster the w_i and re-run the algorithm on each cluster. This is due to issues where we might have, say, 3 clusters, where the first two are relatively close together but very far from the third one. In this case our algorithm would resolve the third cluster from the first two, but needs to be run a second time to then resolve the first two clusters from each other.

Charikar et al. [20] also use this re-clustering idea, but their re-clustering algorithm makes use of a sophisticated metric embedding technique and is relatively complex. Here we avoid this complexity by making use of *resilient sets*, an idea introduced in Steinhardt et al. [61]. A resilient set is a set such that all large subsets have mean close to the mean of the original set; it can be shown that any set with bounded moment tensor is resilient, and by finding such resilient sets we can robustly cluster in a much more direct manner than before. In particular, in the well-separated case we show that after enough rounds of re-clustering, every resilient set has almost all of its points coming from a single cluster, leading to substantially improved error bounds in that case.

REFERENCES

[1] D. Achlioptas and F. McSherry. 2005. On spectral learning of mixtures of distributions. In *Conference on Learning Theory (COLT)*.

[2] R. Adamczak and P. Wolff. 2015. Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields* 162 (2015), 531–586.

[3] Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham Kakade, and Yi-Kai Liu. 2012. A Spectral Algorithm for Latent Dirichlet Allocation. In *NIPS*. 926–934.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. 2013. Tensor decompositions for learning latent variable models. *arXiv* (2013).

[5] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. 2014. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research* 15, 1 (2014), 2773–2832.

[6] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. R. Voss. 2014. The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. In *Conference on Learning Theory (COLT)*.

[7] P. Awasthi and O. Sheffet. 2012. Improved spectral-norm bounds for clustering. *Approximation, Randomization, and Combinatorial Optimization* (2012), 37–49.

[8] M. Balcan, A. Blum, and S. Vempala. 2008. A discriminative framework for clustering via similarity functions. In *Symposium on Theory of Computing (STOC)*.

[9] B. Barak, F. Brandão, A. Harrow, J. Kelner, D. Steurer, and Y. Zhou. 2012. Hypercontractivity, sum-of-squares proofs, and their applications. In *Symposium on Theory of Computing (STOC)*. 307–326.

[10] Boaz Barak, Jonathan A. Kelner, and David Steurer. 2015. Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method. In *STOC*. ACM, 143–151.

[11] Boaz Barak and Ankur Moitra. 2016. Noisy Tensor Completion via the Sum-of-Squares Hierarchy. In *COLT (JMLR Workshop and Conference Proceedings)*, Vol. 49. JMLR.org, 417–445.

[12] M. Belkin and K. Sinha. 2010. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS)*.

[13] T. Bernholt. 2006. *Robust estimators are hard to compute.* Technical Report. Universität Dortmund.

[14] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. 2014. Smoothed analysis of tensor decompositions. In *Symposium on Theory of Computing (STOC)*.

[15] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. 2014. Smoothed analysis of tensor decompositions. In *STOC*. ACM, 594–603.

[16] A. Bhaskara, M. Charikar, and A. Vijayaraghavan. 2014. Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Conference on Learning Theory (COLT)*.

[17] Emmanuel J. Candès and Benjamin Recht. 2009. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics* 9, 6 (2009), 717–772.

[18] Emmanuel J. Candès, Mark Rudelson, Terence Tao, and Roman Vershynin. 2005. Error Correction via Linear Programming. In *FOCS*. IEEE Computer Society, 295–308.

[19] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. 2017. Learning from untrusted data. In *STOC*. ACM, 47–60.

[20] M. Charikar, J. Steinhardt, and G. Valiant. 2017. Learning from Untrusted Data. In *Symposium on Theory of Computing (STOC)*.

[21] Yeshwanth Cherapanamjeri, Prateek Jain, and Praneeth Netrapalli. 2017. Thresholding Based Outlier Robust PCA. In *COLT (Proceedings of Machine Learning Research)*, Vol. 65. PMLR, 593–628.

[22] Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso. 2007. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Trans. Signal Process.* 55, 6, part 2 (2007), 2965–2973. https://doi.org/10.1109/TSP.2007.893943

[23] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. 2016. Robust Estimators in High Dimensions without the Computational Intractability. In *Foundations of Computer Science (FOCS)*.

[24] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. 2017. Being Robust (in High Dimensions) Can Be Practical. *arXiv* (2017).

[25] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2016. Robust Estimators in High Dimensions without the Computational Intractability. In *FOCS*. IEEE Computer Society, 655–664.

[26] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2017. Being Robust (in High Dimensions) Can Be Practical. *CoRR* abs/1703.00893 (2017).

[27] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2017. Robustly Learning a Gaussian: Getting Optimal Error, Efficiently. *CoRR* abs/1704.03866 (2017).

[28] R. Ge, Q. Huang, and S. M. Kakade. 2015. Learning mixtures of Gaussians in high dimensions. In *Symposium on Theory of Computing (STOC)*.

[29] Rong Ge and Tengyu Ma. 2015. Decomposing Overcomplete 3rd Order Tensors using Sum-of-Squares Algorithms. In *APPROX-RANDOM (LIPIcs)*, Vol. 40. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 829–849.

[30] Navin Goyal, Santosh Vempala, and Ying Xiao. 2013. Fourier PCA. *CoRR* abs/1306.5825 (2013).

[31] N. Goyal, S. Vempala, and Y. Xiao. 2014. Fourier PCA and robust tensor decomposition. In *Symposium on Theory of Computing (STOC)*.

[32] David Gross. 2011. Recovering Low-Rank Matrices From Few Coefficients in Any Basis. *IEEE Trans. Information Theory* 57, 3 (2011), 1548–1566.

[33] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 2011. *Robust Statistics: The Approach Based on Influence Functions.* Wiley.

[34] Sam B. Hopkins and Jerry Li. 2017. Mixture Models, Robustness, and Sum of Squares Proofs. (2017).

[35] Samuel B. Hopkins, Jonathan Shi, and David Steurer. 2015. Tensor principal component analysis via sum-of-square proofs. In *COLT (JMLR Workshop and Conference Proceedings)*, Vol. 40. JMLR.org, 956–1006.

[36] D. Hsu and S. M. Kakade. 2013. Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions. In *Innovations in Theoretical Computer Science (ITCS)*.

[37] Daniel Hsu and Sham M. Kakade. 2013. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS'13—Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science*. ACM, New York, 11–19.

[38] D. Hsu, S. M. Kakade, and T. Zhang. 2009. A spectral algorithm for learning hidden Markov models. In *Conference on Learning Theory (COLT)*.

[39] Daniel J. Hsu and Sham M. Kakade. 2013. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *ITCS*. ACM, 11–20.

[40] Peter J Huber. 2011. Robust statistics. In *International Encyclopedia of Statistical Science*. Springer, 1248–1251.

[41] A. T. Kalai, A. Moitra, and G. Valiant. 2010. Efficiently learning mixtures of two Gaussians. In *Symposium on Theory of Computing (STOC)*. 553–562.

[42] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. 2010. Efficiently learning mixtures of two Gaussians. In *STOC*. ACM, 553–562.

[43] P. Kothari and D. Steurer. 2017. Outlier-robust moment-estimation via sum-of-squares. *arXiv* (2017).

[44] Pravesh K. Kothari and Jacob Steinhardt. 2017. Better Agnostic Clustering via Relaxed Tensor Norms. (2017).

[45] A. Kumar and R. Kannan. 2010. Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS)*. 299–308.

[46] K. A. Lai, A. B. Rao, and S. Vempala. 2016. Agnostic Estimation of Mean and Covariance. In *Foundations of Computer Science (FOCS)*.

[47] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. 2016. Agnostic Estimation of Mean and Covariance. In *FOCS*. IEEE Computer Society, 665–674.

[48] R. Latała. 2006. Estimates of moments and tails of Gaussian chaoses. *The Annals of Probability* 34, 6 (2006), 2315–2331.

[49] Tengyu Ma, Jonathan Shi, and David Steurer. 2016. Polynomial-Time Tensor Decompositions with Sum-of-Squares. In *FOCS*. IEEE Computer Society, 438–446.

[50] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. 2014. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability* 42, 3 (2014), 906–945.

[51] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. 2009. The planar k-means problem is NP-hard. *International Workshop on Algorithms and Computation* (2009), 274–285.

[52] R. A. Maronna, D. R. Martin, and V. J. Yohai. 2006. *Robust Statistics: Theory and Methods*. Wiley.

[53] A. Moitra and G. Valiant. 2010. Settling the polynomial learnability of mixtures of Gaussians. In *Foundations of Computer Science (FOCS)*.

[54] Ankur Moitra and Gregory Valiant. 2010. Settling the Polynomial Learnability of Mixtures of Gaussians. In *FOCS*. IEEE Computer Society, 93–102.

[55] Elchanan Mossel and Sébastien Roch. 2005. Learning nonsingular phylogenies and hidden Markov models. In *STOC*. ACM, 366–375.

[56] K. Pearson. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* 185 (1894), 71–110.

[57] Aaron Potechin and David Steurer. 2017. Exact tensor completion with sum-of-squares. In *COLT (Proceedings of Machine Learning Research)*, Vol. 65. PMLR, 1619–1673.

[58] Benjamin Recht. 2011. A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research* 12 (2011), 3413–3430.

[59] O. Regev and A. Vijayaraghavan. 2017. On Learning Mixtures of Well-Separated Gaussians. In *Foundations of Computer Science (FOCS)*.

[60] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. 2017. Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. *CoRR* abs/1703.04940 (2017).

[61] J. Steinhardt, M. Charikar, and G. Valiant. 2018. Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. In *Innovations in Theoretical Computer Science (ITCS)*.

[62] John W. Tukey. 1975. Mathematics and the picturing of data. (1975), 523–531.

[63] S. Vempala and G. Wang. 2002. A spectral algorithm for learning mixture models. In *Foundations of Computer Science (FOCS)*.

[64] Santosh Vempala and Ying Xiao. 2015. Max vs Min: Tensor Decomposition and ICA with nearly Linear Sample Complexity. In *COLT (JMLR Workshop and Conference Proceedings)*, Vol. 40. JMLR.org, 1710–1723.