The Sample Complexity of Robust Covariance Testing

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin-Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the problem of testing the covariance matrix of a high-dimensional Gaussian in a robust setting, where the input distribution has been corrupted in Huber's contamination model. Specifically, we are given i.i.d. samples from a distribution of the form $Z = (1-\epsilon)X + \epsilon B$, where X is a zero-mean and unknown covariance Gaussian $\mathcal{N}(0, \Sigma)$, B is a fixed but unknown noise distribution, and $\epsilon > 0$ is an arbitrarily small constant representing the proportion of contamination. We want to distinguish between the cases that Σ is the identity matrix versus γ-far from the identity in Frobenius norm.

In the absence of contamination, prior work gave a simple tester for this hypothesis testing task that uses $O(d)$ samples. Moreover, this sample upper bound was shown to be best possible, within constant factors. Our main result is that the sample complexity of covariance testing dramatically increases in the contaminated setting. In particular, we prove a sample complexity lower bound of $\Omega(d^2)$ for ϵ an arbitrarily small constant and $\gamma = 1/2$. This lower bound is best possible, as $O(d^2)$ samples suffice to even robustly *learn* the covariance. The conceptual implication of our result is that, for the natural setting we consider, robust hypothesis testing is at least as hard as robust estimation.

Keywords: Robust High-Dimensional Statistics, Hypothesis Testing

1. Introduction

1.1. Background and Motivation

This work can be viewed as a confluence of two research areas: distribution testing and high-dimensional robust statistics. To put our contributions in context, we provide the necessary background.

Distribution Property Testing Distribution property testing Goldreich and Ron (2000); Batu et al. (2000, 2013) is a field at the intersection of property testing Rubinfeld and Sudan (1996); Goldreich et al. (1998) and statistical hypothesis testing Neyman and Pearson (1933); Lehmann and Romano (2005). The standard question in this area is the following: Given sample access to an unknown probability distribution (or, more generally, collection of distributions), how many samples do we need to determine whether the underlying distribution(s) satisfies a pre-specified property or is far, in a well-defined sense, from satisfying the property? This TCS style definition turns out to be essentially equivalent to the minimax view of statistical hypothesis testing, pioneered in mathematical statistics by Ingster and coauthors (see, e.g., Ingster and Suslina (2003).)

During the past few decades, distribution property testing has received significant attention within the computer science and statistics communities. The reader is referred to Rubinfeld (2012); Canonne (2015) for two surveys on the topic. The classical setting typically studied in the relevant

literature concerns testing properties of discrete distributions, where the only available information is an upper bound on the domain size. This setting is fairly well understood. For a range of natural and important properties, there exist testers that require provably minimum sample complexity (up to universal constant factors). See Paninski (2008); Chan et al. (2014); Valiant and Valiant (2014); Diakonikolas et al. (2015a); Acharya et al. (2015); Canonne et al. (2016); Diakonikolas and Kane (2016); Diakonikolas et al. (2016a, 2017a); Canonne et al. (2017b); Neykov et al. (2020); Diakonikolas et al. (2020) for some representative works. A key conceptual message of this line of work is that the *sample complexity of testing* is *significantly lower* than the sample complexity of *learning* the underlying distribution(s).

More recently, a body of work in computer science has focused on leveraging *a priori structure* of the underlying distributions to obtain significantly improved sample complexities, see, e.g., Batu et al. (2004); Daskalakis et al. (2013); Diakonikolas et al. (2015a,b); Canonne et al. (2017a); Daskalakis and Pan (2017); Daskalakis et al. (2018); Diakonikolas et al. (2017b, 2019); Canonne et al. (2019). Specifically, a line of work has established that it is possible to efficiently test various properties of *high-dimensional* structured distributions — including high-dimensional Gaussians, discrete product distributions, and various graphical models — with sample complexity significantly better than learning the distribution. Importantly, these algorithmic results are fragile, in the sense that they crucially rely on thee assumption that the underlying distribution satisfies the given structure *exactly*.

High-Dimensional Robust Statistics Robust statistics is the subfield of statistics focusing on the design of estimators that are *robust* to deviations from the modeling assumptions (see, e.g., Hampel et al. (1986); Huber and Ronchetti (2009) for introductory statistical textbooks on the topic). A learning algorithm is *robust* if its performance is stable to deviations from the idealized assumptions about the input data. The precise form of this deviation depends on the setting and gives rise to various definitions of robustness. Here we focus on *Huber's contamination model* Huber (1964), which prescribes that an adversary generates samples from a mixture distribution P of the form $P = (1-\epsilon)D + \epsilon B$, where D is the unknown target distribution and B is an adversarially chosen noise distribution. The parameter $\epsilon \in [0, 1/2)$ is the proportion of contamination and quantifies the power of the adversary. Intuitively, among our samples, an unknown $(1-\epsilon)$ fraction are generated from a distribution of interest and are called *inliers*, and the rest are called *outliers*.

It is well-known that standard estimators (e.g., the empirical mean) crucially rely on the assumption that the observations are generated from the assumed model (e.g., an unknown Gaussian). The existence of even a *single* outlier can arbitrarily compromise their performance. Classical work in robust statistics developed robust estimators with optimal sample complexity for several basic high-dimensional learning tasks. For example, the Tukey median Tukey (1975) is a sample-efficient robust mean estimator for spherical Gaussian distributions. These early robust estimators are not computationally efficient, in particular they incur runtime exponential in the dimension. More recently, a successful line of work in computer science, starting with Diakonikolas et al. (2016b); Lai et al. (2016), has lead to *computationally efficient* robust learning algorithms in a wide range of high-dimensional settings. The reader is referred to Diakonikolas and Kane (2019) for a recent survey.

This Work In sharp contrast to the sample complexity of robust learning (which is fairly well-understood for several natural settings), the sample complexity of *robust testing* in high dimensions is poorly understood. While various aspects of robust hypothesis testing have been studied in the

robust statistics literature Wilcox (1997), a number of basic questions remain wide-open. A natural research direction is to understand how the robustness requirement affects the complexity of high-dimensional testing in various parametric settings. In particular, if the underlying distribution *nearly* satisfies the assumed structural property (e.g., is *almost* a multivariate Gaussian, as opposed to exactly one) can we still obtain testers with sub-learning sample complexity?

In this paper, we focus on the fundamental problem of *testing the covariance matrix* of a high-dimensional distribution. This is a classical question that has seen renewed interest from the statistics community, see, e.g., Cai and Ma (2013); Cai et al. (2016a,b) and Cai (2017) for an overview article. The most basic problem formulation is the following: We are given n samples from an unknown Gaussian distribution $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^d$ with zero mean and unknown covariance. We want to distinguish, with probability at least $2/3$, between the cases that $\Sigma = I$ versus $\|\Sigma - I\|_F \geq \gamma$, where $\|\cdot\|_F$ denotes Frobenius norm.

In the noiseless setting, this testing question was studied in Cai and Ma (2013); Cai et al. (2016a), where it was shown that $\Theta(d/\gamma^2)$ samples are necessary and sufficient. On the other hand, the sample complexity of learning the covariance within error γ in Frobenius norm is $\Theta(d^2/\gamma^2)$.

In the rejoinder article Cai et al. (2016b) of Cai et al. (2016a), Balasubramanian and Yuan gave a counterexample for the tester proposed of Cai et al. (2016a) in the presence of contamination and explicitly raised the question of understanding the sample complexity of robust covariance testing in Huber's model. They write

> "Much work is still needed to gain fundamental understanding of robust estimation under ϵ-contamination model or other reasonable models."

The robust covariance testing question is the following: We are given n samples from an unknown distribution on $\mathbb{R}^d$ of the form $(1-\epsilon)\mathcal{N}(0, \Sigma) + \epsilon B$, where B is an unknown noise distribution. We want to distinguish, with probability at least $2/3$, between the cases that $\Sigma = I$ versus $\|\Sigma - I\|_F \geq \gamma$. Importantly, for this statistical task to be information-theoretically possible, we need to assume that the contamination fraction ϵ is significantly smaller than the "gap" γ between the completeness and soundness cases.

In summary, we ask the following question:

Question 1.1 *What is the sample complexity of* robustly *testing the covariance matrix of a high-dimensional Gaussian with respect to the Frobenius norm?*

Our main result (Theorem 1) is that for ϵ an arbitrarily small positive constant and $\gamma = 1/2$, robust covariance testing requires $\Omega(d^2)$ samples. This bound is best possible, as with $O(d^2)$ samples we can robustly estimate the covariance matrix within the desired error. This answers the open question posed in Cai et al. (2016b).

In summary, our result shows that there is a *quadratic gap* between the sample complexity of testing and robust testing for the covariance. *Notably, such a sample complexity gap does not exist for the problems of learning and robustly learning the covariance.* In particular, the robust learning version of the problem has the same sample complexity as its non-robust version. That is, the robustness requirement makes the testing problem *information-theoretically* harder – a phenomenon that does *not* appear in the context of learning.

Prior work Diakonikolas et al. (2017c) has shown an analogous phenomenon for the much simpler problem of robustly testing the mean of a Gaussian. As we explain in the following section,

the techniques of Diakonikolas et al. (2017c) inherently fail for the covariance setting, and it seemed plausible that better robust covariance testers could exist.

1.2. Our Results and Techniques

Our main result is the following theorem:

Theorem 1 *For any constants $C, \epsilon > 0$, any algorithm that can distinguish between the standard Gaussian $\mathcal{N}(0, I)$ on $\mathbb{R}^d$ and a distribution $X = (1-\epsilon)\mathcal{N}(0, \Sigma) + \epsilon B$, for some Σ with $\|\Sigma - I\|_F > C$, requires $\Omega(d^2)$ samples.*

Proof Overview and Comparison to Prior Work. Our sample complexity lower bounds will make use of standard information-theoretic tools and, of course, the innovation is in constructing and analyzing the hard instances.

At a very high level, our techniques are similar to those used in Diakonikolas et al. (2017c) that gave a lower bound for robust mean testing. Diakonikolas et al. (2017c) defined an adversarial ensemble $\mathcal{D}$ which was a distribution over noisy Gaussians with means far from 0. They proceeded to show that the distribution $\mathcal{D}^N$ (defined as taking a random distribution from $\mathcal{D}$ and then returning N i.i.d. samples from that distribution) was close in total variation distance to G^N, where G is the standard Gaussian. This was done by bounding the χ^2 distance between the corresponding product distributions, and in particular showing that

$$\chi^2_{G^N}(\mathcal{D}^N, \mathcal{D}^N) = 1 + o(1) \, .$$

To prove this bound, one notes that

$$\chi^2_{G^N}(\mathcal{D}^N, \mathcal{D}^N) = \mathbf{E}_{P,Q\sim\mathcal{D}}\chi^2_{G^n}(P^N, Q^N) = \mathbf{E}_{P,Q\sim\mathcal{D}}(\chi^2_G(P, Q))^N.$$

The desired implication follows if it can be shown that $\chi^2_G(P, Q)$ is close to 1 with high probability.

Such a bound could be proven relatively easily for the mean case, given the techniques developed in Diakonikolas et al. (2017c). In particular, P and Q could be chosen to be distributions that were a standard Gaussian in all but one special direction. And in this direction they were copies of some reference distribution A, which matched its first few moments with a Gaussian. The technology developed in Diakonikolas et al. (2017c) then allowed one to show that $\chi^2_G(P, Q) - 1$ would be small unless the defining directions of these two distributions were close to each other (which in d dimensions happens with very small probability for two random unit vectors).

Importantly, this kind of construction *cannot* be made to work for the robust covariance testing problem. In particular, if the adversarial distributions look like standard Gaussians in all but one direction, this can easily be detected robustly using only $O(d)$ samples for the following reason: Two random vectors in d-dimensions will be close to each other with probability only exponentially small in d. In order to prove an $\Omega(d^2)$ lower bound using these ideas, the defining "directions" for the bad distributions must be drawn from a d^2-dimensional space. Given that the covariance matrix is d^2-dimensional, it is not hard to see what this space might be, however a new analysis is needed because one cannot readily produce a distribution that is a standard Gaussian in all orthogonal directions.

For our new construction, we take A to be a symmetric matrix and let our hard distribution be $\mathcal{N}(0, I + A)$, with noise added as a copy of $\mathcal{N}(0, I - MA)$ for some appropriate constant $M > 0$,

so that the average of the covariance matrices (taking the weights of the components into account) is the identity. We call this distribution $[\widetilde{A}]$. We choose our adversarial ensemble to return $[\widetilde{A}]$ for an appropriately chosen random matrix A. Since we cannot use the technology from Diakonikolas et al. (2017c) to bound the chi-squared distance, we need new technical ideas. Specifically, we are able to obtain a formula for $\chi^2_G([\widetilde{A}], [\widetilde{B}])$. This allows us to show the following: Assuming that A and B have small operator norms (and we can restrict to only using matrices for which this is the case), then we have that $\chi^2_G([\widetilde{A}], [\widetilde{B}])$ is small, so long as $\mathrm{tr}(AB)^2$ and $\mathrm{tr}((AB)^2)$ are. A careful analysis of these polynomials in the coefficients of A and B gives that both are well concentrated for random matrices A and B to make our analysis go through.

2. Proof of Main Result

Throughout this proof, we will treat ϵ and C as constants and will thus suppress dependence on then in our asymptotic notation.

We will require a few definitions and basic facts.

Definition 2 *We begin by recalling the chi-squared inner product. For distributions A, B, C we have that*

$$\chi^2_A(B, C) = \int \frac{dB dC}{dA}.$$

We also note the following elementary fact:

Fact 3 *For distributions A and B we have that $d_{\mathrm{TV}}(A, B) \geq \sqrt{\chi^2_A(B, B) - 1}/2$.*

Proof It is easy to see that $\chi^2_A(B, B) - 1 = \int \frac{(dB - dA)(dB - dA)}{dA}$. By Cauchy-Schwartz, this is bigger than

$$\left(\int \frac{|dB - dA| dA}{dA}\right)^2 / \int \frac{dA dA}{dA} = 4 d_{\mathrm{TV}}(A, B)^2.$$

■

We are now ready to proceed with our proof. We begin by directly computing the formula for the chi-squared distance of two mean zero Gaussians with respect to a third.

Lemma 4 *Let $\Sigma_1, \Sigma_2 \prec 2I$ be positive definite symmetric matrices. Then,*

$$\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0, \Sigma_1), \mathcal{N}(0, \Sigma_2)) = (\det(\Sigma_1 + \Sigma_2 - \Sigma_1 \Sigma_2))^{-1/2}.$$

Proof Letting $p(x), p_1(x), p_2(x)$ be the probability density functions for $\mathcal{N}(0, I)$, $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$, respectively. We then have that

$$\begin{aligned}
&\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0,\Sigma_1),\mathcal{N}(0,\Sigma_2)) \\
&\quad= \int \frac{p_1(x)p_2(x)}{p(x)}dx \\
&\quad= \int \frac{(2\pi)^{-d/2}(\det(\Sigma_1))^{-1/2}\exp(-x^T\Sigma_1^{-1}x/2)(2\pi)^{-d/2}(\det(\Sigma_2))^{-1/2}\exp(-x^T\Sigma_2^{-1}x/2)}{(2\pi)^{-d/2}\exp(-x^Tx/2)}dx \\
&\quad= (\det(\Sigma_1\Sigma_2))^{-1/2}\int (2\pi)^{-d/2}\exp(-x^T(\Sigma_1^{-1}+\Sigma_2^{-1}-I)x/2)dx \\
&\quad= (\det(\Sigma_1\Sigma_2))^{-1/2}(\det(\Sigma_1^{-1}+\Sigma_2^{-1}-I))^{-1/2} \\
&\quad= (\det(\Sigma_1+\Sigma_2-\Sigma_1\Sigma_2))^{-1/2}\ .
\end{aligned}$$

This completes the proof. ■

We can then approximate this quantity using Taylor expansion. The result is particularly nice if the covariances are $I + A$ and $I + B$, where A and B have small operator norms.

Lemma 5 *If A, B are symmetric matrices with $\|A\|_2, \|B\|_2 = O(1/\sqrt{d})$, then*

$$\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0, I+A), \mathcal{N}(0, I+B)) = (1 + \mathrm{tr}(AB)/2 + O(\mathrm{tr}(AB)^2 + \mathrm{tr}((AB)^2) + 1/d^2)).$$

Proof Applying Lemma 4 gives that the term in question is

$$(\det((I+A)+(I+B)-(I+A)(I+B)))^{-1/2} = (\det(I-AB))^{-1/2}.$$

Suppose that AB has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then

$$\begin{aligned}
\det(I-AB) = \prod_{i=1}^n (1-\lambda_i) &= \exp(-\sum_{i=1}^n \sum_{m=1}^\infty \lambda_i^m/m + O(\lambda_i^2)) \\
&= \exp(-\sum_{m=1}^\infty \sum_{i=1}^n \lambda_i^m/m + O(\lambda_i^2)) \\
&= \exp(-\sum_{m=1}^\infty \mathrm{tr}((AB)^m)/m).
\end{aligned}$$

We note that $\mathrm{tr}((AB)^m) = O(1/d)^{m-1}$. Thus, this expression is

$$\exp(-\mathrm{tr}(AB))(1 - \mathrm{tr}((AB)^2)/2)(1 + O(1/d^2)).$$

Therefore,

$$\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0, I+A), \mathcal{N}(0, I+B)) = (1 + \mathrm{tr}(AB)/2 + O(\mathrm{tr}(AB)^2 + \mathrm{tr}((AB)^2) + 1/d^2))\ .$$

This proves our lemma. ■

The key noisy Gaussians that will show up in our adversarial ensemble will be of the following form:

Definition 6 *Let A be a symmetric matrix. Define the probability distribution $\widetilde{[A]}$ as follows:*

$$\widetilde{[A]} = (1-\epsilon)\mathcal{N}(0, I + 2CA) + \epsilon\mathcal{N}(0, I - (2C(1-\epsilon)/\epsilon)A) \ .$$

Notice that this is carefully chosen so that the average covariance of these two components is exactly the identity.

Next we need to compute the chi-squared inner product of two of these $\widetilde{[A]}$ distributions with respect to the standard Gaussian. This is not hard as we already know the inner products of the Gaussian components.

Lemma 7 *For A and B matrices with $\|A\|_2, \|B\|_2 = O(1/\sqrt{d})$, we have that*

$$\chi^2_{\mathcal{N}(0,I)}(\widetilde{[A]}, \widetilde{[B]}) = 1 + O(((\mathrm{tr}(AB)^2 + \mathrm{tr}((AB)^2) + 1/d^2)).$$

Proof Letting $C' = C(1-\epsilon)/\epsilon$, we have that $\chi^2_{\mathcal{N}(0,I)}(\widetilde{[A]}, \widetilde{[B]})$ equals

$$(1-\epsilon)^2\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0, I+2CA), \mathcal{N}(0, I+B) + \epsilon(1-\epsilon)\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0, I+2CA), \mathcal{N}(0, I-2C'B))$$
$$+\epsilon(1-\epsilon)\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0, I-2C'A), \mathcal{N}(0, I+2CB) + \epsilon^2\chi^2_{\mathcal{N}(0,I)}(\mathcal{N}(0, I-2C'A), \mathcal{N}(0, I-2C'B)).$$

We expand each term using Lemma 5 and note that the $\mathrm{tr}(AB)$ terms all cancel. The remaining terms are as desired. ■

We can now define our adversarial ensemble that will be hard to distinguish from a standard Gaussian.

Definition 8 *Let $\mathcal{D}$ be the following ensemble. Pick a symmetric matrix A whose diagonal entries are 0 and whose off-diagonal entries are $\mathcal{N}(0, 1/d)$ random variables, independent except for the symmetry, all conditioned on $\|A\|_2 = O(1/\sqrt{d})$ and $\|A\|_F = O(1)$ (note that these both happen with high probability). Let $\mathcal{D}$ return the distribution $\widetilde{[A]}$. Note that with high probability $\widetilde{[A]}$ is of the form $(1-\epsilon)\mathcal{N}(0, \Sigma) + \epsilon B$ for some distribution B and some Σ with $\|\Sigma - I\|_F > C$.*

Let $\mathcal{D}^N$ denote the distribution over $\mathbb{R}^{d\times N}$ obtained by picking a random distribution X from $\mathcal{D}$ and taking N i.i.d. samples from it.

Let $G = \mathcal{N}(0, I)$ and G^N denote the distribution obtained by taking N i.i.d. samples from G.

We prove the following crucial proposition:

Proposition 9 *For $N = o(d^2)$, we have that*

$$\chi^2_{G^N}(\mathcal{D}^N, \mathcal{D}^N) = 1 + o(1) \ .$$

Proof We begin by noting that

$$\chi^2_{G^N}(\mathcal{D}^N, \mathcal{D}^N) = \mathbf{E}_{\widetilde{[A]},\widetilde{[B]}\sim\mathcal{D}}[\chi^2_{G^N}(\widetilde{[A]}^N, \widetilde{[B]}^N)] = \mathbf{E}_{\widetilde{[A]},\widetilde{[B]}\sim\mathcal{D}}[\chi^2_G(\widetilde{[A]}, \widetilde{[B]})^N].$$

Applying Lemma 7, this is

$$\mathbf{E}_{\widetilde{[A]},\widetilde{[B]}\sim\mathcal{D}}[(1 + O(\mathrm{tr}(AB)^2 + \mathrm{tr}((AB)^2) + O(1/d^2)))^N].$$

In order to bound this quantity, we wish to show that $\mathrm{tr}(AB)^2$ and $\mathrm{tr}((AB)^2)$ are both small with high probability. We note that A and B are both random Gaussian symmetric matrices conditioned on some high probability event C.

We first note that fixing A that without conditioning on C we have that $\mathrm{tr}(AB)$ is distributed as a normal random variable with standard deviation $O(\|A\|_F/d) = O(1/d)$. Therefore, $\mathrm{tr}(AB)^2 \leq t$ except with probability $\exp(-\Omega(d^2 t))$.

Next we wish to similarly understand the distribution of $\mathrm{tr}((AB)^2)$. We note that

$$\mathrm{tr}((AB)^2) = \sum_{i,j,k,\ell} A_{ij} B_{jk} A_{k\ell} B_{\ell i}$$

is a quadratic polynomial in each of A and B. All of the terms except for those with $\{i,j\} = \{k,\ell\}$ have mean 0, and the remaining terms have mean $O(1/d^2)$. Thinking of B as fixed, $A_{ij}A_{k\ell}$ has coefficient $B_{jk}B_{\ell i} + B_{ik}B_{j\ell}$. Thus, as a polynomial in A, the variance is $O(\|B\|_F^2/d^2) = O(1/d^2)$. Therefore, by standard concentration inequalities, the probability that $\mathrm{tr}((AB)^2) > t$ is $\exp(-\Omega(d^2 t))$.

Hence, we have shown that $\mathrm{tr}(AB)^2 + \mathrm{tr}((AB)^2) > t$ with probability $\exp(-\Omega(d^2 t))$. In particular, it is stochastically dominated by $O(\mathcal{N}(0,1)^2/d^2)$.

Back to our original problem, we wish to bound

$$\mathbf{E}_{\widetilde{[A]},\widetilde{[B]}\sim\mathcal{D}}[(1 + O(\mathrm{tr}(AB)^2 + \mathrm{tr}((AB)^2)) + O(1/d^2))^N].$$

By the above, this is less than

$$\mathbf{E}[(1 + O(\mathcal{N}(0,1)^2/d^2) + O(1/d^2))^N] \leq \mathbf{E}[\exp(O(N/d^2)\mathcal{N}(0,1)^2 + O(N/d^2))] = 1 + o(1)\ ,$$

which completes the proof. ■

We can now prove our main theorem:

Theorem 10 *For any $C, \epsilon > 0$, there is no algorithm that distinguishes between $\mathcal{N}(0, I)$ and a distribution X obtained by adding ϵ additive noise to $\mathcal{N}(0, \Sigma)$ for some Σ with $\|\Sigma - I\|_F > C$, using $N = o(d^2)$ samples.*

Proof We note that such an algorithm could reliably distinguish between a sample from G^N and $\mathcal{D}^N$, which contradicts our proposition. ■

Acknowledgments

Ilias Diakonikolas was supported by NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Daniel M. Kane was supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship.

References

J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *Proceedings of NIPS'15*, 2015.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000. URL citeseer.ist.psu.edu/batu00testing.html.

T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

T. T. Cai. Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annual Review of Statistics and Its Application*, 4(1):423–446, 2017. doi: 10.1146/annurev-statistics-060116-053754.

T. T. Cai and Z. Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388, 2013.

T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Statist.*, 10(1):1–59, 2016a. URL https://doi.org/10.1214/15-EJS1081.

T. T. Cai, Z. Ren, and H. H. Zhou. Rejoinder of ?estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation? *Electron. J. Statist.*, 10(1):81–89, 2016b. doi: 10.1214/15-EJS1081REJ. URL https://doi.org/10.1214/15-EJS1081REJ.

C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, pages 25:1–25:14, 2016. See also **?** (full version).

C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing Bayesian networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 370–448, 2017a.

C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing conditional independence of discrete distributions. *CoRR*, abs/1711.11560, 2017b. URL http://arxiv.org/abs/1711.11560. In STOC'18.

C. L. Canonne, X. Chen, G. Kamath, A. Levi, and E. Waingarten. Random restrictions of high-dimensional distributions and uniformity testing with subcube conditioning. *Electronic Colloquium on Computational Complexity (ECCC)*, 26:165, 2019. URL https://eccc.weizmann.ac.il/report/2019/165.

S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

C. Daskalakis and Q. Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 697–703, 2017.

C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

C. Daskalakis, N. Dikkala, and G. Kamath. Testing Ising models. In *SODA*, 2018. To appear.

I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016. Full version available at abs/1601.05557.

I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019. URL http://arxiv.org/abs/1911.05911.

I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, 2015a.

I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*, 2015b.

I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Collision-based testers are optimal for uniformity and closeness. *Electronic Colloquium on Computational Complexity (ECCC)*, 23: 178, 2016a.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, pages 655–664, 2016b.

I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Sample-optimal identity testing with high probability. *CoRR*, abs/1708.02728, 2017a. In ICALP 2018.

I. Diakonikolas, D. M. Kane, and V. Nikishkin. Near-optimal closeness testing of discrete histogram distributions. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017*, pages 8:1–8:15, 2017b.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *FOCS*, pages 73–84, 2017c.

I. Diakonikolas, D. M. Kane, and J. Peebles. Testing identity of multidimensional histograms. In *Conference on Learning Theory, COLT 2019*, pages 1107–1131, 2019.

I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price. Optimal testing of discrete distributions with high probability. *CoRR*, abs/2009.06540, 2020. URL https://arxiv.org/abs/2009.06540.

O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000.

O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45:653–750, 1998.

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

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

P.J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 2009.

Y. Ingster and I. A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer, 2003.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Proceedings of FOCS'16*, 2016.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

M. Neykov, S. Balakrishnan, and L. Wasserman. Minimax optimal conditional independence testing. *CoRR*, 2020.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL http://rsta.royalsocietypublishing.org/content/231/694-706/289.short.

L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. on Comput.*, 25:252–271, 1996.

J. W. Tukey. Mathematics and picturing of data. In *Proceedings of ICM*, volume 6, pages 523–531, 1975.

G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

R. R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Statistical modeling and decision science. Acad. Press, San Diego, Calif. [u.a.], 1997.