Robust learning of high-dimensional Gaussian distributions

Bogdan Grechuk • 30 Jun 2026

The Gödel Prize is an annual prize for outstanding papers in theoretical computer science. The 2026 Gödel Prize has been awarded to Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart for their paper (Diakonikolas et al. 2019), which resolves a longstanding problem in robust statistics.

In many modern applications, the data are not just noisy in the usual random sense: a small part of the data may be completely unreliable. A sensor may fail, a database may contain malicious entries, or a very large experiment may include a few measurements produced under the wrong conditions. The classical estimators for the mean and the covariance of a Gaussian distribution are the empirical mean and the empirical covariance. They are excellent when all samples are honest, but they are fragile: a tiny fraction of carefully chosen outliers can move the empirical mean by a large amount in high dimensions.

We now describe a theorem of Diakonikolas, Kamath, Kane, Li, Moitra and Stewart (Diakonikolas et al. 2019), which shows that this obstacle can be overcome algorithmically for Gaussian distributions. We first introduce the precise language needed to state the result.

For a vector $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, let x^T denote its transpose. A real symmetric $d \times d$ matrix A is *positive semidefinite* s]positive semidefinite matrix if $x^T A x \geq 0$ for every $x \in \mathbb{R}^d$, and it is *positive definite* if $x^T A x > 0$ for every non-zero $x \in \mathbb{R}^d$. If $X = (X_1, \dots, X_d)$ is a random vector in $\mathbb{R}^d$, its *mean vector* is

$$\mu = \mathbb{E}X = (\mathbb{E}X_1, \dots, \mathbb{E}X_d),$$

and its *covariance matrix* is the $d \times d$ matrix Σ with entries

$$\Sigma_{ij} = \mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)\big], \quad 1 \leq i, j \leq d.$$

Thus Σ_{ii} is the variance of the i-th coordinate, while Σ_{ij} for $i \neq j$ measures how the i-th and j-th coordinates vary together.

For $\mu \in \mathbb{R}^d$ and a positive definite matrix Σ, the *Gaussian distribution* $\mathcal{N}(\mu, \Sigma)$ is the probability distribution on $\mathbb{R}^d$ with density

$$f(x) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right),$$

where $\det \Sigma$ is the determinant of Σ. The vector μ is the mean vector of the distribution, and the matrix Σ is its covariance matrix.

To measure the quality of an estimated distribution, we use *total variation distance*. If P and Q are two probability distributions on $\mathbb{R}^d$, define

$$d_{\mathrm{TV}}(P, Q) = \sup_A |P(A) - Q(A)|,$$

where the supremum is over all Borel sets $A \subset \mathbb{R}^d$, that is, over all sets whose probabilities are defined from open sets by countable unions, intersections, and complements. Equivalently, if P and Q have densities p and q, then

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)|\, dx.$$

This number lies between 0 and 1. It is 0 exactly when the two distributions are the same, and it controls the advantage of any statistical test trying to distinguish a sample from P from a sample from Q.

Finally, an *ϵ-corrupted set of N samples* from a distribution P is a list obtained as follows. First, N independent samples are drawn from P. Then an adversary, after seeing all the samples, is allowed to replace at most ϵN of them by arbitrary points of $\mathbb{R}^d$. The algorithm receives only the final corrupted list, not the identity of the corrupted points.

We use the asymptotic notation $\widetilde{\Omega}(F)$ to denote $CF \log^c(F+2)$ for some universal constants $C, c > 0$.

Theorem 1 *Let $d \geq 1$, let $0 < \epsilon < 1/2$, and let $0 < \tau < 1$. Let $\mathcal{N}(\mu, \Sigma)$ be an unknown Gaussian distribution on $\mathbb{R}^d$, where both the mean vector μ and the positive definite covariance matrix Σ are unknown. There is a polynomial-time algorithm which, given ϵ, τ, and an ϵ-corrupted set of N samples from $\mathcal{N}(\mu, \Sigma)$ with*

$$N \geq \widetilde{\Omega}\left(\frac{d^2 \log^5(1/\tau)}{\epsilon^2}\right),$$

outputs a vector $\widehat{\mu} \in \mathbb{R}^d$ and a positive definite matrix $\widehat{\Sigma}$ such that, with probability at least $1 - \tau$,

$$d_{\mathrm{TV}}\big(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})\big) \leq O\left(\epsilon \log^{3/2} \frac{1}{\epsilon}\right).$$

The most striking feature of Theorem 1 is that the final error bound has is independent of the dimension d. Of course, the number of required samples must grow with d: even without corruptions, estimating an arbitrary covariance matrix requires learning about $d(d+1)/2$ numbers. But once enough samples are available, the theorem says that the effect of the adversary is controlled essentially by the corruption rate ϵ, not by the ambient dimension.

In the same work (Diakonikolas et al. 2019), the authors also proved versions of Theorem 1 for many other important families of distributions, including a product distribution on the hypercube, mixtures of two product distributions, and mixtures of spherical Gaussians. These results became a turning point in robust statistics. Classical robust estimators in high dimensions often had good mathematical guarantees but were computationally intractable. This theorem shows, for some of the central families of probability distributions, that robustness and efficient computation can coexist: even when an adversary corrupts a constant fraction of the data, a polynomial-time algorithm can still recover a distribution that is almost indistinguishable from the truth.

References

Diakonikolas, Ilias, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2019. “Robust Estimators in High-Dimensions Without the Computational Intractability.” *SIAM J. Comput.* 48 (2): 742–864. https://doi.org/10.1137/17M1126680.