

9 Robust Multivariate Mean Estimation

Before considering corruptions, let's recall what is possible with clean i.i.d. samples $X_1, \dots, X_n \sim \mathcal{D}$ in $\mathbb{R}^d$. Write $\mu = \mathbb{E}[X_i]$ and assume isotropic covariance $\Sigma = I_d$. Here I_d denote the $d \times d$ identity matrix. The isotropic assumption is the same as saying that each coordinate of X_i is sampled independently from $\mathcal{N}(0, 1)$. For an isotropic Gaussian $\mathcal{D} = \mathcal{N}(\mu, I_d)$, the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ satisfies

$$\bar{X} \sim \mathcal{N}(\mu, \frac{1}{n} I_d), \quad \mathbb{E}[\|\bar{X} - \mu\|_2^2] = \frac{d}{n}.$$

Gaussian concentration also gives, with probability at least $1 - \delta$,

$$\|\bar{X} - \mu\|_2^2 = O\left(\frac{d + \log(1/\delta)}{n}\right).$$

These bounds on the error are both optimal in the sense that no estimator can improve either bound on the error, so the sample mean is the best we can do for clean data.

If we assume only $\Sigma = I_d$, with no assumptions on higher moments, the expected squared error of the sample mean is still exactly d/n . The empirical mean will no longer give the high-probability bound that we get for Gaussian distributions, although there are other estimators that achieve the same high-probability statement. We now ask how these estimation guarantees change when an adversary can corrupt a fraction of the observations.

9.1 Things that Don't Work

For both examples, take clean data from $\mathcal{N}(0, I_d)$ and let the adversary put a τ fraction of observations at $a = (1, \dots, 1)$, where $0 < \tau \leq 1/4$. We ignore sampling error to focus on the effect of corruption.

Estimating each coordinate separately. A natural idea is to take the median of each coordinate. The corrupted points put a τ fraction of mass at 1 in every coordinate, shifting every median by $\Theta(\tau)$. These errors add up in Euclidean norm:

$$\|\hat{\mu} - \mu\|_2 = \sqrt{\sum_{j=1}^d \Theta(\tau^2)} = \Theta(\tau\sqrt{d}).$$

Discarding points with large norm. Another idea is to discard points with unusually large norm, then average the rest. Gaussian points typically have norm of order $\sqrt{d}$, so a cutoff such as $2\sqrt{d}$ retains most clean points and also every corrupted point a . The surviving clean points have mean zero by symmetry. The corrupted points make up a $\Theta(\tau)$ fraction of the remaining data and all point in the same direction, giving

$$\|\hat{\mu} - \mu\|_2 = \Theta(\tau)\|a\|_2 = \Theta(\tau\sqrt{d}).$$

9.2 Resilience as a Paradigm for Robustness

Definition 9.1 (Resilience [SCV18]). A sample $S = (x_1, \dots, x_m)$ is (ρ, η) -resilient around μ if every subset $T \subseteq [m]$ with $|T| \geq (1 - \eta)m$ satisfies

$$\|\bar{x}_T - \mu\| \leq \rho, \quad \bar{x}_T = \frac{1}{|T|} \sum_{i \in T} x_i.$$

Here $\rho \geq 0$ and $0 \leq \eta < 1$. The center μ can be the sample mean or the population mean. Note that the norm $\|\cdot\|$ can be any norm we want in this definition.

Resilience obviously gives robustness to deletions: after deleting at most an η fraction of the sample, the mean of the remaining points is still within ρ of μ . More surprisingly, we can also handle insertions, although the algorithm we use becomes much more computationally expensive.

Theorem 9.2 (Robustness to arbitrary insertions [SCV18]). Let $\alpha \geq 0$ and $0 \leq \tau < 1/2$, and set

$$\eta = \frac{\tau}{1 - \tau}.$$

Suppose a clean sample S of $(1 - \tau)n$ points is $(\alpha/2, \eta)$ -resilient around μ in the norm $\|\cdot\|$. Let $\tilde{S}$ be obtained by inserting τn arbitrary points into S . There is an estimator using only $\tilde{S}$ and the parameters α, τ that outputs $\hat{\mu}$ satisfying

$$\|\hat{\mu} - \mu\| \leq \alpha$$

for every choice of the inserted points.

Proof sketch. Search for any subset $S' \subseteq \tilde{S}$ of size $(1 - \tau)n$ that is $(\alpha/2, \eta)$ -resilient around some center $\hat{\mu}$, and output $\hat{\mu}$. Such a subset exists because S itself qualifies. All subsets here refer to sample indices, so repeated values are allowed.

The intersection $T = S \cap S'$ satisfies

$$|T| \geq |S| + |S'| - n = (1 - 2\tau)n = (1 - \eta)|S| = (1 - \eta)|S'|.$$

Thus T is large enough for resilience of both sets to apply. Its mean is within $\alpha/2$ of both centers, so

$$\|\hat{\mu} - \mu\| \leq \|\hat{\mu} - \bar{x}_T\| + \|\bar{x}_T - \mu\| \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha. \quad \square$$

Gaussian tails give particularly strong resilience in the Euclidean norm. The following statement includes the sampling error relative to the population mean.

Lemma 9.3 (Gaussian resilience in the Euclidean norm). *Let $X_1, \dots, X_m$ be i.i.d. samples from $\mathcal{N}(\mu, I_d)$, and fix $0 < \eta \leq 1/2$ and $\delta \in (0, 1/2)$. With probability at least $1 - \delta$, the sample $S = (X_1, \dots, X_m)$ is (ρ, η) -resilient around μ in $\|\cdot\|_2$, where*

$$\rho = O\left(\eta\sqrt{\log(1/\eta)} + \sqrt{\frac{d + \log(1/\delta)}{m}}\right).$$

Equivalently, every subset $T \subseteq [m]$ with $|T| \geq (1 - \eta)m$ simultaneously satisfies $\|\bar{X}_T - \mu\|_2 \leq \rho$.

Proof sketch. Write $Y_i = X_i - \mu$, so for any fixed nonempty subset $A \subseteq [m]$, the sum $\sum_{i \in A} Y_i$ has distribution $\mathcal{N}(0, |A|I_d)$. Gaussian concentration therefore gives, for $t > 0$,

$$\mathbb{P}\left[\left\|\sum_{i \in A} Y_i\right\|_2 > \sqrt{|A|}(\sqrt{d} + \sqrt{2t})\right] \leq e^{-t}.$$

There are

$$N = \sum_{k=0}^{\lfloor \eta m \rfloor} \binom{m}{k} \leq \exp(m\eta \log(e/\eta))$$

possible deleted subsets. Apply the tail bound with $t = \log(2/\delta)$ to $A = [m]$ and with $t = \log(2N/\delta)$ to each nonempty deleted subset. A union bound makes the total failure probability at most $\delta/2 + N \cdot \delta/(2N) = \delta$; the empty subset has sum zero. Using $|R| \leq \eta m$ and the bound on N gives, simultaneously with probability at least $1 - \delta$,

$$\begin{aligned} \left\|\frac{1}{m} \sum_{i=1}^m Y_i\right\|_2 &\leq O\left(\sqrt{\frac{d + \log(1/\delta)}{m}}\right), \\ \max_{R: |R| \leq \eta m} \frac{1}{m} \left\|\sum_{i \in R} Y_i\right\|_2 &\leq O\left(\eta\sqrt{\log(1/\eta)} + \sqrt{\frac{\eta(d + \log(1/\delta))}{m}}\right). \end{aligned}$$

On this event, let T be any retained subset and put $R = [m] \setminus T$. Since $|T| \geq (1 - \eta)m \geq m/2$,

$$\|\bar{X}_T - \mu\|_2 \leq \frac{m}{|T|} \left(\left\|\frac{1}{m} \sum_{i=1}^m Y_i\right\|_2 + \frac{1}{m} \left\|\sum_{i \in R} Y_i\right\|_2 \right) = O\left(\eta\sqrt{\log(1/\eta)} + \sqrt{\frac{d + \log(1/\delta)}{m}}\right).$$

This holds simultaneously for all such T , with probability at least $1 - \delta$. $\square$

Combining this with Theorem 9.2, using $m = (1 - \tau)n$ and $\eta = \tau/(1 - \tau)$, gives an estimator under arbitrary insertions with Euclidean error

$$\|\hat{\mu} - \mu\|_2 = O\left(\tau\sqrt{\log(1/\tau)} + \sqrt{\frac{d + \log(1/\delta)}{n}}\right)$$

with probability at least $1 - \delta$.

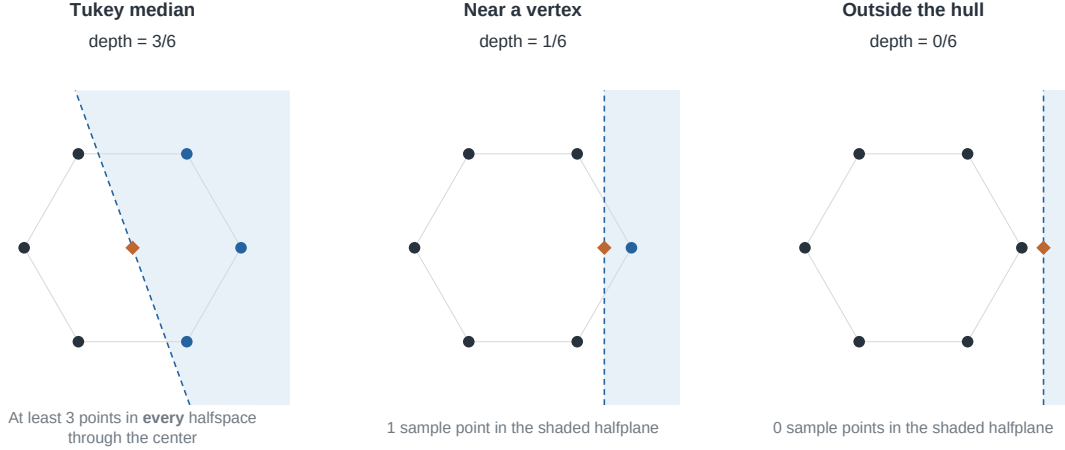


Figure 4: Tukey depth for six points at the vertices of a regular hexagon. Orange diamonds mark the query points; each shaded halfspace attains the minimum count for its query point.

9.3 The Tukey Median

In one dimension, a median has at least half the observations on either side. In higher dimensions, we can ask how well a point balances the sample in every direction. This leads to Tukey depth, also called *halfspace depth*.

Definition 9.4 (Tukey depth). For a sample $X = (X_1, \dots, X_n)$ in $\mathbb{R}^d$, the Tukey depth of a point $z \in \mathbb{R}^d$ is

$$\text{depth}(z; X) = \min_{v \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\langle v, X_i - z \rangle \geq 0\}.$$

An equivalent definition of Tukey depth is that it is the smallest fraction of the sample in any closed halfspace containing z . Another equivalent definition is that it is the minimum fraction of points you have to remove so that z is outside the convex hull of the remaining points.

For a distribution $\mathcal{D}$, define Tukey depth by

$$\text{depth}(z; \mathcal{D}) = \inf_{v \in \mathbb{R}^d: \|v\|_2=1} \mathbb{P}[\langle v, Y - z \rangle \geq 0].$$

A *Tukey median* is any point of maximum depth:

$$\mu_{\text{Tukey}}(X) \in \arg \max_{z \in \mathbb{R}^d} \text{depth}(z; X).$$

It need not be a sample point or be unique. In one dimension this recovers the usual median, including the interval of medians between the two middle observations when n is even.

Depth for Gaussians. For $\mathcal{D} = \mathcal{N}(\mu, I_d)$, every projection onto a unit vector is a one-dimensional Gaussian with variance 1. The halfspace with the least mass points away from μ , giving

$$\text{depth}(z; \mathcal{D}) = \Phi(-\|z - \mu\|_2),$$

where Φ is the standard Gaussian cumulative distribution function. Thus μ has depth $1/2$, and points far from μ have small depth.

Theorem 9.5 (Robustness of the Tukey median for Gaussians). *Let $X_1, \dots, X_n$ be i.i.d. samples from $N(\mu, I_d)$. There is a universal constant $c > 0$ such that, if*

$$\tau + \sqrt{\frac{d + \log(1/\delta)}{n}} \leq c,$$

then with probability at least $1 - \delta$ over the clean sample, simultaneously for every sample $\tilde{X}$ obtained by replacing at most τn of the dataset, and every Tukey median of $\tilde{X}$,

$$\left\| \mu_{\text{Tukey}}(\tilde{X}) - \mu \right\|_2^2 = O\left(\tau^2 + \frac{d + \log(1/\delta)}{n}\right).$$

Proof sketch. The family of halfspaces in $\mathbb{R}^d$ has VC dimension $d + 1$. Uniform convergence therefore bounds the difference between empirical and population halfspace probabilities by

$$\beta = O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)$$

with probability at least $1 - \delta$. Replacing τn points changes the fraction in any halfspace by at most τ . Hence, uniformly in z ,

$$\left| \text{depth}(z; \tilde{X}) - \text{depth}(z; \mathcal{D}) \right| \leq \beta + \tau.$$

Write $\hat{\mu} = \mu_{\text{Tukey}}(\tilde{X})$. Since $\hat{\mu}$ maximizes empirical depth and the population depth of μ is $1/2$,

$$\text{depth}(\hat{\mu}; \mathcal{D}) \geq \text{depth}(\hat{\mu}; \tilde{X}) - (\beta + \tau) \geq \text{depth}(\mu; \tilde{X}) - (\beta + \tau) \geq \frac{1}{2} - 2(\beta + \tau).$$

The Gaussian depth formula then implies

$$\|\hat{\mu} - \mu\|_2 \leq \Phi^{-1}\left(\frac{1}{2} + 2(\beta + \tau)\right) = O(\beta + \tau)$$

when $\beta + \tau$ is sufficiently small. □

The corruption-dependent error has no dependence on d ; dimension enters through sampling error. Computing a Tukey median in high dimensions, however, is computationally expensive, motivating the efficient algorithms we discuss next.

9.4 Efficient Algorithms

The examples in Section 9.1 show that corrupted points can shift the mean while looking harmless in each coordinate and having ordinary norms. Their effect becomes visible when we examine the covariance matrix.

For a clean sample $X_1, \dots, X_n \sim \mathcal{N}(0, I_d)$, the empirical covariance

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top$$

is close to I_d when n is sufficiently large compared with d . In particular, the variance in every unit direction v is approximately $v^\top \widehat{\Sigma} v \approx 1$.

Fact 9.6 (Gaussian covariance concentration). *Let $X_1, \dots, X_n$ be i.i.d. samples from $\mathcal{N}(\mu, I_d)$, and let $\widehat{\Sigma}$ be their empirical covariance as defined above. There is a universal constant $C > 0$ such that, for any $\varepsilon, \delta \in (0, 1)$, if*

$$n \geq C \frac{d + \log(1/\delta)}{\varepsilon^2},$$

then with probability at least $1 - \delta$, simultaneously for every unit vector $v \in \mathbb{R}^d$,

$$1 - \varepsilon \leq v^\top \widehat{\Sigma} v \leq 1 + \varepsilon.$$

Equivalently, every eigenvalue of $\widehat{\Sigma}$ lies in $[1 - \varepsilon, 1 + \varepsilon]$.

Now replace a τ fraction of the observations by $a = (1, \dots, 1)$. Ignoring sampling error, the corrupted distribution has mean $\tilde{\mu} = \tau a$ and covariance

$$\widetilde{\Sigma} = (1 - \tau)I_d + \tau(1 - \tau)aa^\top.$$

The first term comes from the clean points' covariance; the second comes from the separation between the clean and corrupted means. Along $v = a/\sqrt{d}$, the variance is

$$v^\top \widetilde{\Sigma} v = (1 - \tau)(1 + \tau d),$$

while in every direction perpendicular to a it is $1 - \tau$. Thus a $\tau\sqrt{d}$ shift in the mean produces a covariance eigenvalue of order $1 + \tau d$.

A large eigenvalue therefore gives us both a way to detect this corruption and a direction in which to look for the responsible points. This is the starting point for efficient filtering algorithms for robust mean estimation [Dia+16].

9.4.1 The Filtering Algorithm

Suppose the clean distribution is $\mathcal{N}(\mu, I_d)$ and a τ fraction of the sample has been corrupted. The filtering algorithm repeatedly checks whether the current sample has too much variance in some direction. If it does, we use that direction to reduce the influence of the corrupted points.

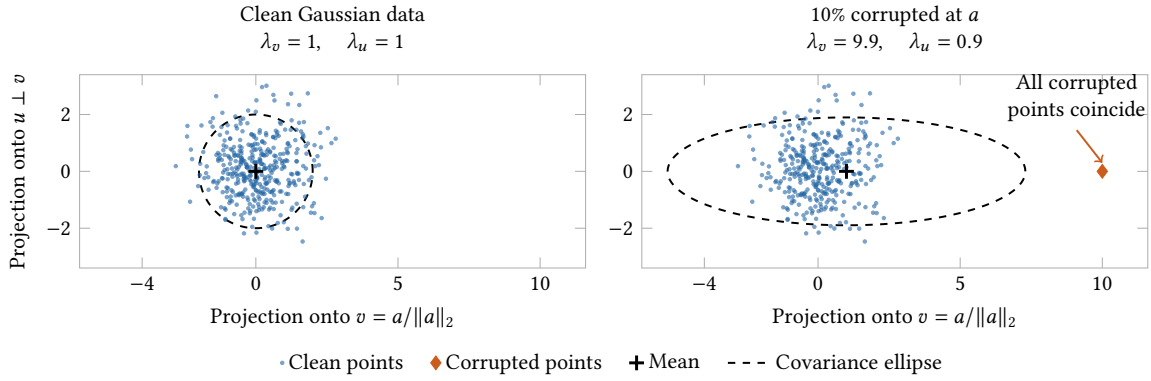


Figure 5: Covariance before and after corruption, with $d = 100$ and $\tau = 0.1$. We project onto the direction $v = a/\sqrt{d}$ and an orthogonal unit direction u . Dots illustrate Gaussian observations; the orange diamond represents all corrupted points, which project to $(\sqrt{d}, 0)$. Dashed ellipses have radii twice the population standard deviations and are centered at the respective means. The variance along v grows from 1 to 9.9, although every corrupted point has norm $\sqrt{d}$ in the original space.

Algorithm 5: Filtering for Gaussian mean estimation (high level)

Start with the corrupted sample S . Repeat the following steps:

1. Compute the mean $\bar{X}_S$ and covariance $\widehat{\Sigma}_S$ of the current sample.
2. If the largest eigenvalue of $\widehat{\Sigma}_S$ is sufficiently close to 1, return $\bar{X}_S$.
3. Otherwise, find a unit eigenvector v corresponding to the largest eigenvalue and assign each point x_i the score

$$s_i = \langle v, x_i - \bar{X}_S \rangle^2.$$

4. Use Gaussian tail bounds to choose a filtering rule that removes or downweights points with large scores. Update S and repeat.

If points are downweighted, use the weighted mean and covariance in subsequent iterations.

Theorem 9.7 (Gaussian filtering guarantee [Dia+16]). *There are universal constants $c, C > 0$ such that the following holds. Let $0 < \tau < c$, $\delta \in (0, 1/2)$, and*

$$n \geq C \frac{d + \log(1/\delta)}{\tau^2}.$$

Draw n independent samples from $\mathcal{N}(\mu, I_d)$, and let an adversary replace at most τn of them arbitrarily. With suitable choices of the stopping threshold and filtering rule, the filtering algorithm runs in time $\text{poly}(n, d)$ and returns the mean $\tilde{\mu}$ of the final filtered sample, satisfying

$$\|\tilde{\mu} - \mu\|_2 \leq C\tau\sqrt{\log(1/\tau)}.$$

The guarantee holds with probability at least $1 - \delta$ over the clean sample, simultaneously for every allowed corruption. The algorithm can be implemented using at most n filtering iterations.

The key to the stopping rule is that a large shift in the mean forces excess variance in the direction of that shift. For the following calculation, ignore sampling error and idealize the clean covariance as exactly I_d .

Lemma 9.8 (Mean shift forces excess variance). *Suppose the clean points have sample mean $\hat{\mu}$ and covariance I_d . Insert arbitrary points making up a $\tau \in (0, 1/2)$ fraction of the final sample. Let $\tilde{\mu}$ and $\tilde{\Sigma}$ be the mean and covariance of this corrupted sample. Then in the direction $v = (\tilde{\mu} - \hat{\mu}) / \|\tilde{\mu} - \hat{\mu}\|_2$,*

$$v^\top \tilde{\Sigma} v \geq 1 - \tau + \frac{1 - \tau}{\tau} \|\tilde{\mu} - \hat{\mu}\|_2^2.$$

Proof. Let b and Σ_B be the mean and covariance of the inserted points. Write $\Delta = \tilde{\mu} - \hat{\mu}$. The mean identity $\tilde{\mu} = (1 - \tau)\hat{\mu} + \tau b$ gives $b - \hat{\mu} = \Delta/\tau$. The covariance identity then gives

$$\tilde{\Sigma} = (1 - \tau)I_d + \tau\Sigma_B + \frac{1 - \tau}{\tau} \Delta\Delta^\top.$$

Since Σ_B is positive semidefinite, taking the variance along $v = \Delta / \|\Delta\|_2$ proves the claim. $\square$