Supplementary Material for Being Robust (in High Dimensions) Can Be Practical

Ilias Diakonikolas
CS, USC
diakonik@usc.edu

Gautam Kamath
EECS & CSAIL, MIT
g@csail.mit.edu

Daniel M. Kane
CSE & Math, UCSD
dakane@cs.ucsd.edu

Jerry Li
EECS & CSAIL, MIT
jerryzli@mit.edu

Ankur Moitra
Math & CSAIL, MIT
moitra@mit.edu

Alistair Stewart
CS, USC
alistais@usc.edu

1 Omitted Details from Section 3

1.1 Robust Mean Estimation for Sub-Gaussian Distributions

In this section, we use our filter technique to give a *near sample-optimal* computationally efficient algorithm to robustly estimate the mean of a sub-gaussian density with a known covariance matrix, thus proving Theorem 3.1.

We emphasize that the algorithm and its analysis is essentially identical to the filtering algorithm given in Section 8.1 of [DKK+16] for the case of a Gaussian $\mathcal{N}(\mu, I)$. The only difference is a weaker definition of the "good set of samples" (Definition 2) and a simple concentration argument (Lemma 1) showing that a random set of uncorrupted samples of the appropriate size is good with high probability. Given these, the analysis of this subsection follows straightforwardly from the analysis in Section 8.1 of [DKK+16] by plugging in the modified parameters. For the sake of completeness, we provide the details below.

We start by formally defining sub-gaussian distributions:

Definition 1. *A distribution P on R with mean μ, is sub-gaussian with parameter $\nu > 0$ if*

$$\mathbb{E}_{X\sim P}\left[\exp(\lambda(X-\mu))\right] \leq \exp(\nu\lambda^2/2)$$

for all $\lambda \in \mathsf{R}$. A distribution P on R^d with mean vector μ is sub-gaussian with parameter $\nu > 0$, if for all unit vectors v, the one-dimensional random variable $v \cdot X$, $X \sim P$, is sub-gaussian with parameter ν.

We will use the following simple fact about the concentration of sub-gaussian random variables:

Fact 1. *If P is sub-gaussian on R^d with mean vector μ and parameter $\nu > 0$, then for any unit vector $v \in \mathsf{R}^d$ we have that $\Pr_{X\sim P}\left[|v \cdot (X-\mu)| \geq T\right] \leq \exp(-t^2/2\nu)$.*

The following theorem is a high probability version of Theorem 3.1:

Theorem 2. *Let G be a sub-gaussian distribution on R^d with parameter $\nu = \Theta(1)$, mean μ^G, covariance matrix I, and $\varepsilon, \tau > 0$. Let S' be an ε-corrupted set of samples from G of size $\Omega((d/\varepsilon^2)\mathrm{poly}\log(d/\varepsilon\tau))$. There exists an efficient algorithm that, on input S' and $\varepsilon > 0$, returns a mean vector $\widehat{\mu}$ so that with probability at least $1-\tau$ we have $\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

Notation. We will denote $\mu^S = \frac{1}{|S|}\sum_{X\in S} X$ and $M_S = \frac{1}{|S|}\sum_{X\in S}(X-\mu^G)(X-\mu^G)^T$ for the sample mean and modified sample covariance matrix of the set S.

We start by defining our modified notion of good sample, i.e, a set of conditions on the uncorrupted set of samples under which our algorithm will succeed.

Definition 2. *Let G be an identity covariance sub-gaussian in d dimensions with mean μ^G and covariance matrix I and $\varepsilon, \tau > 0$. We say that a multiset S of elements in R^d is* (ε,τ)-good with respect to G *if the following conditions are satisfied:*

(i) For all $x \in S$ we have $\|x-\mu^G\|_2 \le O(\sqrt{d\log(|S|/\tau)})$.

(ii) For every affine function $L : \mathsf{R}^d \to \mathsf{R}$ such that $L(x) = v\cdot(x-\mu^G) - T$, $\|v\|_2 = 1$, we have that $|\Pr_{X\in_u S}[L(X)\ge 0] - \Pr_{X\sim G}[L(X)\ge 0]| \le \frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}$.

(iii) We have that $\|\mu^S - \mu^G\|_2 \le \varepsilon$.

(iv) We have that $\|M_S - I\|_2 \le \varepsilon$.

We show in the following subsection that a sufficiently large set of independent samples from G is (ε,τ)-good (with respect to G) with high probability. Specifically, we prove:

Lemma 1. *Let G be sub-gaussian distribution with parameter $\nu = \Theta(1)$ and with identity covariance, and $\varepsilon, \tau > 0$. If the multiset S is obtained by taking $\Omega((d/\varepsilon^2)\mathrm{poly}\log(d/\varepsilon\tau))$ independent samples from G, it is (ε,τ)-good with respect to G with probability at least $1-\tau$.*

We require the following definition that quantifies the extent to which a multiset has been corrupted:

Definition 3. *Given finite multisets S and S' we let $\Delta(S,S')$ be the size of the symmetric difference of S and S' divided by the cardinality of S.*

The starting point of our algorithm will be a simple NaivePrune routine (Section 4.3.1 of [DKK+16]) that removes obvious outliers, i.e., points which are far from the mean. Then, we iterate the algorithm whose performance guarantee is given by the following:

Proposition 1. *Let G be a sub-gaussian distribution on R^d with parameter $\nu = \Theta(1)$, mean μ^G, covariance matrix I, $\varepsilon > 0$ be sufficiently small and $\tau > 0$. Let S be an (ε,τ)-good set with respect to G. Let S' be any multiset with $\Delta(S,S') \le 2\varepsilon$ and for any $x, y \in S'$, $\|x-y\|_2 \le O(\sqrt{d\log(d/\varepsilon\tau)})$. There exists a polynomial time algorithm* Filter-Sub-Gaussian-Unknown-Mean *that, given S' and $\varepsilon > 0$, returns one of the following:*

(i) A mean vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

(ii) A multiset $S'' \subseteq S'$ such that $\Delta(S,S'') \le \Delta(S,S') - \varepsilon/\alpha$, where $\alpha \stackrel{\mathrm{def}}{=} d\log(d/\varepsilon\tau)\log(d\log(\frac{d}{\varepsilon\tau}))$.

We start by showing how Theorem 2 follows easily from Proposition 1.

Proof of Theorem 2. By the definition of $\Delta(S,S')$, since S' has been obtained from S by corrupting an ε-fraction of the points in S, we have that $\Delta(S,S') \le 2\varepsilon$. By Lemma 1, the set S of uncorrupted samples is (ε,τ)-good with respect to G with probability at least $1-\tau$. We henceforth condition on this event.

Since S is (ε,τ)-good, all $x \in S$ have $\|x-\mu^G\|_2 \le O(\sqrt{d\log|S|/\tau})$. Thus, the NaivePrune procedure does not remove from S' any member of S. Hence, its output, S'', has $\Delta(S,S'') \le \Delta(S,S')$

and for any $x \in S''$, there is a $y \in S$ with $\|x - y\|_2 \leq O(\sqrt{d \log |S|/\tau})$. By the triangle inequality, for any $x, z \in S''$, $\|x - z\|_2 \leq O(\sqrt{d \log |S|/\tau}) = O(\sqrt{d \log(d/\varepsilon\tau)})$.

Then, we iteratively apply the Filter-Sub-Gaussian-Unknown-Mean procedure of Proposition 1 until it terminates returning a mean vector μ with $\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. We claim that we need at most $O(\alpha)$ iterations for this to happen. Indeed, the sequence of iterations results in a sequence of sets S'_i, so that $\Delta(S, S'_i) \leq \Delta(S, S') - i \cdot \varepsilon/\alpha$. Thus, if we do not output the empirical mean in the first 2α iterations, in the next iteration there are no outliers left and the algorithm terminates outputting the sample mean of the remaining set. □

1.1.1 Algorithm Filter-Sub-Gaussian-Unknown-Mean: Proof of Proposition 1

In this subsection, we describe the efficient algorithm establishing Proposition 1 and prove its correctness. Our algorithm calculates the empirical mean vector $\mu^{S'}$ and empirical covariance matrix Σ. If the matrix Σ has no large eigenvalues, it returns $\mu^{S'}$. Otherwise, it uses the eigenvector v^* corresponding to the maximum magnitude eigenvalue of Σ and the mean vector $\mu^{S'}$ to define a filter. Our efficient filtering procedure is presented in detailed pseudocode below.

Algorithm 1 Filter algorithm for a sub-gaussian with unknown mean and identity covariance

1: **procedure** Filter-Sub-Gaussian-Unknown-Mean(S', ε, τ)
input: A multiset S' such that there exists an (ε, τ)-good S with $\Delta(S, S') \leq 2\varepsilon$
output: Multiset S'' or mean vector $\widehat{\mu}$ satisfying Proposition 1
2: Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance matrix Σ , i.e., $\Sigma = (\Sigma_{i,j})_{1 \leq i,j \leq d}$ with $\Sigma_{i,j} = \mathbb{E}_{X \in_u S'}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$.
3: Compute approximations for the largest absolute eigenvalue of $\Sigma - I$, $\lambda^* := \|\Sigma - I\|_2$, and the associated unit eigenvector v^*.
4: **if** $\|\Sigma - I\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **then return** $\mu^{S'}$.
5: **end if**
6: Let $\delta := 3\sqrt{\varepsilon \|\Sigma - I\|_2}$. Find $T > 0$ such that

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta\right] > 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{T^2 \log\left(d \log(\frac{d}{\varepsilon\tau})\right)}.$$

7: **return** the multiset $S'' = \{x \in S' : |v^* \cdot (x - \mu^{S'})| \leq T + \delta\}$.
8: **end procedure**

1.1.2 Proof of Correctness of Filter-Sub-Gaussian-Unknown-Mean

By definition, there exist disjoint multisets L, E, of points in R^d, where $L \subset S$, such that $S' = (S \setminus L) \cup E$. With this notation, we can write $\Delta(S, S') = \frac{|L|+|E|}{|S|}$. Our assumption $\Delta(S, S') \leq 2\varepsilon$ is equivalent to $|L| + |E| \leq 2\varepsilon \cdot |S|$, and the definition of S' directly implies that $(1 - 2\varepsilon)|S| \leq |S'| \leq (1 + 2\varepsilon)|S|$. Throughout the proof, we assume that ε is a sufficiently small constant.

We define $\mu^G, \mu^S, \mu^{S'}, \mu^L$, and μ^E to be the means of G, S, S', L, and E, respectively.

Our analysis will make essential use of the following matrices:

- $M_{S'}$ denotes $\mathbb{E}_{X \in_u S'}[(X - \mu^G)(X - \mu^G)^T]$,
- M_S denotes $\mathbb{E}_{X \in_u S}[(X - \mu^G)(X - \mu^G)^T]$,
- M_L denotes $\mathbb{E}_{X \in_u L}[(X - \mu^G)(X - \mu^G)^T]$, and

- M_E denotes $\mathbb{E}_{X\in_u E}[(X-\mu^G)(X-\mu^G)^T]$.

Our analysis will hinge on proving the important claim that $\Sigma - I$ is approximately $(|E|/|S'|)M_E$. This means two things for us. First, it means that if the positive errors align in some direction (causing M_E to have a large eigenvalue), there will be a large eigenvalue in $\Sigma - I$. Second, it says that any large eigenvalue of $\Sigma - I$ will correspond to an eigenvalue of M_E, which will give an explicit direction in which many error points are far from the empirical mean.

Useful Structural Lemmas. We begin by noting that we have concentration bounds on G and therefore, on S due to its goodness.

Fact 3. *Let $w \in \mathsf{R}^d$ be any unit vector, then for any $T > 0$, $\Pr_{X\sim G}\left[|w\cdot(X-\mu^G)| > T\right] \leq 2\exp(-T^2/2\nu)$ and $\Pr_{X\in_u S}\left[|w\cdot(X-\mu^G)| > T\right] \leq 2\exp(-T^2/2\nu) + \frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}$.*

Proof. The first line is Fact 1, and the former follows from it using the goodness of S. □

By using the above fact, we obtain the following simple claim:

Claim 1. *Let $w \in \mathsf{R}^d$ be any unit vector, then for any $T > 0$, we have that:*

$$\Pr_{X\sim G}[|w\cdot(X-\mu^{S'})| > T + \|\mu^{S'}-\mu^G\|_2] \leq 2\exp(-T^2/2\nu).$$

and

$$\Pr_{X\in_u S}[|w\cdot(X-\mu^{S'})| > T + \|\mu^{S'}-\mu^G\|_2] \leq 2\exp(-T^2/2\nu) + \frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}.$$

Proof. This follows from Fact 3 upon noting that $|w\cdot(X-\mu^{S'})| > T + \|\mu^{S'}-\mu^G\|_2$ only if $|w\cdot(X-\mu^G)| > T$. □

We can use the above facts to prove concentration bounds for L. In particular, we have the following lemma:

Lemma 2. *We have that $\|M_L\|_2 = O\left(\log(|S|/|L|) + \varepsilon|S|/|L|\right)$.*

Proof. Since $L \subseteq S$, for any $x \in \mathsf{R}^d$, we have that

$$|S|\cdot\Pr_{X\in_u S}(X=x) \geq |L|\cdot\Pr_{X\in_u L}(X=x)\,. \tag{1}$$

Since M_L is a symmetric matrix, we have $\|M_L\|_2 = \max_{\|v\|_2=1}|v^T M_L v|$. So, to bound $\|M_L\|_2$ it suffices to bound $|v^T M_L v|$ for unit vectors v. By definition of M_L, for any $v \in \mathsf{R}^d$ we have that

$$|v^T M_L v| = \mathbb{E}_{X\in_u L}[|v\cdot(X-\mu^G)|^2].$$

For unit vectors v, the RHS is bounded from above as follows:

$$\begin{aligned}
\mathbb{E}_{X\in_u L}\left[|v\cdot(X-\mu^G)|^2\right] &= 2\int_0^\infty \Pr_{X\in_u L}\left[|v\cdot(X-\mu^G)|>T\right]TdT \\
&= 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \Pr_{X\in_u L}[|v\cdot(X-\mu^G)|>T]TdT \\
&\le 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \min\left\{1,\frac{|S|}{|L|}\cdot\Pr_{X\in_u S}\left[|v\cdot(X-\mu^G)|>T\right]\right\}TdT \\
&\ll \int_0^{4\sqrt{\nu\log(|S|/|L|)}} TdT \\
&+ (|S|/|L|)\int_{4\sqrt{\nu\log(|S|/|L|)}}^{O(\sqrt{d\log(d/\varepsilon\tau)})}\Big(\exp(-T^2/2\nu)+\frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}\Big)TdT \\
&\ll \log(|S|/|L|)+\varepsilon\cdot|S|/|L|\ ,
\end{aligned}$$

where the second line follows from the fact that $\|v\|_2=1$, $L\subset S$, and S satisfies condition (i) of Definition 2; the third line follows from (1); and the fourth line follows from Fact 3. □

As a corollary, we can relate the matrices $M_{S'}$ and M_E, in spectral norm:

Corollary 1. *We have that $M_{S'}-I=(|E|/|S'|)M_E+O(\varepsilon\log(1/\varepsilon))$, where the $O(\varepsilon\log(1/\varepsilon))$ term denotes a matrix of spectral norm $O(\varepsilon\log(1/\varepsilon))$.*

Proof. By definition, we have that $|S'|M_{S'}=|S|M_S-|L|M_L+|E|M_E$. Thus, we can write

$$\begin{aligned}
M_{S'} &= (|S|/|S'|)M_S-(|L|/|S'|)M_L+(|E|/|S'|)M_E \\
&= I+O(\varepsilon)+O(\varepsilon\log(1/\varepsilon))+(|E|/|S'|)M_E\ ,
\end{aligned}$$

where the second line uses the fact that $1-2\varepsilon\le|S|/|S'|\le1+2\varepsilon$, the goodness of S (condition (iv) in Definition 2), and Lemma 2. Specifically, Lemma 2 implies that $(|L|/|S'|)\|M_L\|_2=O(\varepsilon\log(1/\varepsilon))$. Therefore, we have that

$$M_{S'}=I+(|E|/|S'|)M_E+O(\varepsilon\log(1/\varepsilon))\ ,$$

as desired. □

We now establish a similarly useful bound on the difference between the mean vectors:

Lemma 3. *We have that $\mu^{S'}-\mu^G=(|E|/|S'|)(\mu^E-\mu^G)+O(\varepsilon\sqrt{\log(1/\varepsilon)})$, where the $O(\varepsilon\sqrt{\log(1/\varepsilon)})$ term denotes a vector with ℓ_2-norm at most $O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

Proof. By definition, we have that

$$|S'|(\mu^{S'}-\mu^G)=|S|(\mu^S-\mu^G)-|L|(\mu^L-\mu^G)+|E|(\mu^E-\mu^G).$$

Since S is a good set, by condition (iii) of Definition 2, we have $\|\mu^S-\mu^G\|_2=O(\varepsilon)$. Since $1-2\varepsilon\le|S|/|S'|\le1+2\varepsilon$, it follows that $(|S|/|S'|)\|\mu^S-\mu^G\|_2=O(\varepsilon)$. Using the valid inequality $\|M_L\|_2\ge\|\mu^L-\mu^G\|_2^2$ and Lemma 2, we obtain that $\|\mu^L-\mu^G\|_2\le O\left(\sqrt{\log(|S|/|L|)}+\sqrt{\varepsilon|S|/|L|}\right)$. Therefore,

$$(|L|/|S'|)\|\mu^L-\mu^G\|_2\le O\left((|L|/|S|)\sqrt{\log(|S|/|L|)}+\sqrt{\varepsilon|L|/|S|}\right)=O(\varepsilon\sqrt{\log(1/\varepsilon)})\ .$$

In summary,

$$\mu^{S'} - \mu^G = (|E|/|S'|)(\mu^E - \mu^G) + O(\varepsilon\sqrt{\log(1/\varepsilon)}) \ ,$$

as desired. This completes the proof of the lemma. □

By combining the above, we can conclude that $\Sigma - I$ is approximately proportional to M_E. More formally, we obtain the following corollary:

Corollary 2. *We have $\Sigma - I = (|E|/|S'|)M_E + O(\varepsilon \log(1/\varepsilon)) + O(|E|/|S'|)^2\|M_E\|_2$, where the additive terms denote matrices of appropriately bounded spectral norm.*

Proof. By definition, we can write $\Sigma - I = M_{S'} - I - (\mu^{S'} - \mu^G)(\mu^{S'} - \mu^G)^T$. Using Corollary 1 and Lemma 3, we obtain:

$$\begin{aligned}\Sigma - I &= (|E|/|S'|)M_E + O(\varepsilon \log(1/\varepsilon)) + O((|E|/|S'|)^2\|\mu^E - \mu^G\|_2^2) + O(\varepsilon^2 \log(1/\varepsilon)) \\ &= (|E|/|S'|)M_E + O(\varepsilon \log(1/\varepsilon)) + O(|E|/|S'|)^2\|M_E\|_2 \ ,\end{aligned}$$

where the second line follows from the valid inequality $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$. This completes the proof. □

Case of Small Spectral Norm. We are now ready to analyze the case that the mean vector $\mu^{S'}$ is returned by the algorithm in Step 4. In this case, we have that $\lambda^* \stackrel{\text{def}}{=} \|\Sigma - I\|_2 = O(\varepsilon \log(1/\varepsilon))$. Hence, Corollary 2 yields that

$$(|E|/|S'|)\|M_E\|_2 \leq \lambda^* + O(\varepsilon \log(1/\varepsilon)) + O(|E|/|S'|)^2\|M_E\|_2 \ ,$$

which in turns implies that

$$(|E|/|S'|)\|M_E\|_2 = O(\varepsilon \log(1/\varepsilon)) \ .$$

On the other hand, since $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$, Lemma 3 gives that

$$\|\mu^{S'} - \mu^G\|_2 \leq (|E|/|S'|)\sqrt{\|M_E\|_2} + O(\varepsilon\sqrt{\log(1/\varepsilon)}) = O(\varepsilon\sqrt{\log(1/\varepsilon)}).$$

This proves part (i) of Proposition 1.

Case of Large Spectral Norm. We next show the correctness of the algorithm when it returns a filter in Step 6.

We start by proving that if $\lambda^* \stackrel{\text{def}}{=} \|\Sigma - I\|_2 > C\varepsilon \log(1/\varepsilon)$, for a sufficiently large universal constant C, then a value T satisfying the condition in Step 6 exists. We first note that that $\|M_E\|_2$ is appropriately large. Indeed, by Corollary 2 and the assumption that $\lambda^* > C\varepsilon \log(1/\varepsilon)$ we deduce that

$$(|E|/|S'|)\|M_E\|_2 = \Omega(\lambda^*) \ . \tag{2}$$

Moreover, using the inequality $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$ and Lemma 3 as above, we get that

$$\|\mu^{S'} - \mu^G\|_2 \leq (|E|/|S'|)\sqrt{\|M_E\|_2} + O(\varepsilon\sqrt{\log(1/\varepsilon)}) \leq \delta/2 \ , \tag{3}$$

where we used the fact that $\delta \stackrel{\text{def}}{=} \sqrt{\varepsilon\lambda^*} > C'\varepsilon\sqrt{\log(1/\varepsilon)}$.

Suppose for the sake of contradiction that for all $T > 0$ we have that

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta\right] \leq 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{T^2 \log\left(d\log(\frac{d}{\varepsilon\tau})\right)} \ .$$

Using (3), we obtain that for all $T > 0$ we have that

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^G)| > T + \delta/2\right] \leq 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{T^2 \log\left(d\log(\frac{d}{\varepsilon\tau})\right)} \,. \tag{4}$$

Since $E \subseteq S'$, for all $x \in \mathsf{R}^d$ we have that $|S'|\Pr_{X\in_u S'}[X = x] \geq |E|\Pr_{Y\in_u E}[Y = x]$. This fact combined with (4) implies that for all $T > 0$

$$\Pr_{X \in_u E}\left[|v^* \cdot (X - \mu^G)| > T + \delta/2\right] \ll (|S'|/|E|)\left(\exp(-T^2/2\nu) + \frac{\varepsilon}{T^2 \log\left(d\log(\frac{d}{\varepsilon\tau})\right)}\right) \,. \tag{5}$$

We now have the following sequence of inequalities:

$$\begin{aligned}
\|M_E\|_2 &= \mathbb{E}_{X\in_u E}\left[|v^* \cdot (X - \mu^G)|^2\right] = 2\int_0^\infty \Pr_{X\in_u E}\left[|v^* \cdot (X - \mu^G)| > T\right] T dT \\
&= 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \Pr_{X\in_u E}\left[|v^* \cdot (X - \mu^G)| > T\right] T dT \\
&\leq 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \min\left\{1, \frac{|S'|}{|E|}\cdot \Pr_{X\in_u S'}\left[|v^* \cdot (X - \mu^G)| > T\right]\right\} T dT \\
&\ll \int_0^{4\sqrt{\nu\log(|S'|/|E|)}+\delta} T dT + (|S'|/|E|)\int_{4\sqrt{\nu\log(|S'|/|E|)}+\delta}^{O(\sqrt{d\log(d/\varepsilon\tau)})} \Big(\exp(-T^2/2\nu) + \frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}\Big) T dT \\
&\ll \log(|S'|/|E|) + \delta^2 + O(1) + \varepsilon\cdot|S'|/|E| \\
&\ll \log(|S'|/|E|) + \varepsilon\lambda^* + \varepsilon\cdot|S'|/|E| \,.
\end{aligned}$$

Rearranging the above, we get that

$$(|E|/|S'|)\|M_E\|_2 \ll (|E|/|S'|)\log(|S'|/|E|) + (|E|/|S'|)\varepsilon\lambda^* + \varepsilon = O(\varepsilon\log(1/\varepsilon) + \varepsilon^2\lambda^*).$$

Combined with (2), we obtain $\lambda^* = O(\varepsilon\log(1/\varepsilon))$, which is a contradiction if C is sufficiently large. Therefore, it must be the case that for some value of T the condition in Step 6 is satisfied.

The following claim completes the proof:

Claim 2. *Fix $\alpha \stackrel{\text{def}}{=} d\log(d/\varepsilon\tau)\log(d\log(\frac{d}{\varepsilon\tau}))$. We have that $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/\alpha$.*

Proof. Recall that $S' = (S \setminus L) \cup E$, with E and L disjoint multisets such that $L \subset S$. We can similarly write $S'' = (S \setminus L') \cup E'$, with $L' \supseteq L$ and $E' \subset E$. Since

$$\Delta(S, S') - \Delta(S, S'') = \frac{|E \setminus E'| - |L' \setminus L|}{|S|},$$

it suffices to show that $|E \setminus E'| \geq |L' \setminus L| + \varepsilon|S|/\alpha$. Note that $|L' \setminus L|$ is the number of points rejected by the filter that lie in $S \cap S'$. Note that the fraction of elements of S that are removed to produce S'' (i.e., satisfy $|v^* \cdot (x - \mu^{S'})| > T + \delta$) is at most $2\exp(-T^2/2\nu) + \varepsilon/\alpha$. This follows from Claim 1 and the fact that $T = O(\sqrt{d\log(d/\varepsilon\tau)})$.

Hence, it holds that $|L' \setminus L| \leq (2\exp(-T^2/2\nu) + \varepsilon/\alpha)|S|$. On the other hand, Step 6 of the algorithm ensures that the fraction of elements of S' that are rejected by the filter is at least

$8\exp(-T^2/2\nu) + 8\varepsilon/\alpha)$. Note that $|E \setminus E'|$ is the number of points rejected by the filter that lie in $S' \setminus S$. Therefore, we can write:

$$\begin{aligned}
|E \setminus E'| &\geq (8\exp(-T^2/2\nu) + 8\varepsilon/\alpha)|S'| - (2\exp(-T^2/2\nu) + \varepsilon/\alpha)|S| \\
&\geq (8\exp(-T^2/2\nu) + 8\varepsilon/\alpha)|S|/2 - (2\exp(-T^2/2\nu) + \varepsilon/\alpha)|S| \\
&\geq (2\exp(-T^2/2\nu) + 3\varepsilon/\alpha)|S| \\
&\geq |L' \setminus L| + 2\varepsilon|S|/\alpha \ ,
\end{aligned}$$

where the second line uses the fact that $|S'| \geq |S|/2$ and the last line uses the fact that $|L' \setminus L|/|S| \leq 2\exp(-T^2/2\nu) + \varepsilon/\alpha$. Noting that $\log(d/\varepsilon\tau) \geq 1$, this completes the proof of the claim. □

1.1.3 Proof of Lemma 1

Proof. Let $N = \Omega((d/\varepsilon^2)\mathrm{poly}\log(d/\varepsilon\tau))$ be the number of samples drawn from G. For (i), the probability that a coordinate of a sample is at least $\sqrt{2\nu\log(Nd/3\tau)}$ is at most $\tau/3dN$ by Fact 1. By a union bound, the probability that all coordinates of all samples are smaller than $\sqrt{2\nu\log(Nd/3\tau)}$ is at least $1-\tau/3$. In this case, $\|x\|_2 \leq \sqrt{2\nu d\log(Nd/3\tau)} = O(\sqrt{d\nu\log(N\nu/\tau)})$.

After translating by μ^G, we note that (iii) follows immediately from Lemmas 4.3 of [DKK+16] and (iv) follows from Theorem 5.50 of [Ver10], as long as $N = \Omega(\nu^4 d\log(1/\tau)/\varepsilon^2)$, with probability at least $1-\tau/3$. It remains to show that, conditioned on (i), (ii) holds with probability at least $1-\tau/3$.

To simplify some expressions, let $\delta := \varepsilon/(\log(d\log d/\varepsilon\tau))$ and $R = C\sqrt{d\log(|S|/\tau)}$. We need to show that for all unit vectors v and all $0 \leq T \leq R$ that

$$\left|\Pr_{X\in_u S}[|v\cdot(X-\mu^G)| > T] - \Pr_{X\sim G}[|v\cdot(X-\mu^G) > T \geq 0]\right| \leq \frac{\delta}{T^2} \ . \tag{6}$$

Firstly, we show that for all unit vectors v and $T > 0$

$$\left|\Pr_{X\in_u S}[|v\cdot(X-\mu^G)| > T] - \Pr_{X\sim G}[|v\cdot(X-\mu^G)| > T \geq 0]\right| \leq \delta/4\nu\ln(1/\delta)$$

with probability at least $1-\tau/6$. Since the VC-dimension of the set of all halfspaces is $d+1$, this follows from the VC inequality [DL01], since we have more than $\Omega(d/(\delta/(4\nu\log(1/\delta))^2)$ samples. We thus only need to consider the case when $T \geq \sqrt{4\nu\ln(1/\delta)}$.

Lemma 4. *For any fixed unit vector v and $T > \sqrt{4\nu\ln(1/\delta)}$, except with probability $\exp(-N\delta/6C\nu)$, we have that*

$$\Pr_{X\in_u S}[|v\cdot(X-\mu^G)| > T] \leq \frac{\delta}{CT^2} \ ,$$

where $C = 8$.

Proof. Let E be the event that $|v\cdot(X-\mu^G)| > T$. Since G is sub-gaussian, Fact 1 yields that $\Pr_G[E] = \Pr_{Y\sim G}[|v\cdot(X-\mu^G)| \leq \exp(-T^2/2\nu)$. Note that, thanks to our assumption on T, we have that $T \leq \exp(T^2/4\nu)/2C$, and therefore $T^2\Pr_G[E] \leq \exp(-T^2/4\nu)/2C \leq \delta/2C$.

Consider $\mathbb{E}_S[\exp(t^2/3\nu \cdot N \Pr_S[E])]$. Each individual sample sample X_i for $1 \leq i \leq N$, is an independent copy of $Y \sim G$, and hence:

$$\begin{aligned}
\mathbb{E}_S\left[\exp(T^2/3\nu \cdot N \Pr_S[E])\right] &= \mathbb{E}_S\left[\exp(T^2/3\nu \cdot \sum_{i=1}^{n} 1_{X_i \in E})\right] \\
&= \prod_{i=1}^{N} \mathbb{E}_{X_i}\left[\exp(T^2/3\nu \cdot \sum_{i=1}^{n} 1_{X_i \in E})\right] \\
&= \prod_{i=1}^{N} \exp(T^2/3\nu \cdot \Pr_G[E]) \\
&\leq \exp\left(NT^2/3\nu \cdot \exp(-T^2/2\nu)\right) \\
&\leq \exp(N/3\nu \cdot \delta/2C) = \exp(N\delta/6C\nu) \; .
\end{aligned}$$

Note that if $\Pr_S[E] > \frac{\delta}{C\nu T^2}$, then $\exp(T^2/3\nu \cdot N \Pr_S[E]) = \exp(N\delta/3C\nu)$. By Markov's inequality, this happens with probability at most $\exp(-N\delta/6C\nu)$. □

Now let $\mathcal{C}$ be a 1/2-cover in Euclidean distance for the set of unit vectors of size $2^{O(d)}$. By a union bound, for all $v' \in \mathcal{C}$ and T' a power of 2 between $\sqrt{4\nu \ln(1/\delta)}$ and R, we have that

$$\Pr_{X \in_u S}[|v' \cdot (X - \mu^G)| > T'] \leq \frac{\delta}{8T^2}$$

except with probability

$$2^{O(d)} \log(R) \exp(-N\delta/6C\nu) = \exp\left(O(d) + \log\log R - N\delta/6C\nu\right) \leq \tau/6 \; .$$

However, for any unit vector v and $\sqrt{4\nu \ln(1/\delta)} \leq T \leq R$, there is a $v' \in \mathcal{C}$ and such a T' such that for all $x \in \mathsf{R}^d$, we have $|v \cdot (X - \mu^G)| \geq |v' \cdot (X - \mu^G)|/2$, and so $|v' \cdot (X - \mu^G)| > 2T'$ implies $|v' \cdot (X - \mu^G)| > T$.

Then, by a union bound, (6) holds simultaneously for all unit vectors v and all $0 \leq T \leq R$, with probability a least $1 - \tau/3$. This completes the proof. □

1.2 Robust Mean Estimation Under Second Moment Assumptions

In this section, we use our filtering technique to give a near sample-optimal computationally efficient algorithm to robustly estimate the mean of a density with a second moment assumption. We show:

Theorem 4. *Let P be a distribution on R^d with unknown mean vector μ^P and unknown covariance matrix $\Sigma_P \preceq I$. Let S be an ε-corrupted set of samples from P of size $\Theta((d/\varepsilon)\log d)$. Then there exists an algorithm that given S, with probability $2/3$, outputs $\widehat{\mu}$ with $\|\widehat{\mu} - \mu^P\|_2 \leq O(\sqrt{\varepsilon})$ in time* $\mathrm{poly}(d/\varepsilon)$.

Note that Theorem 3.2 follows straightforwardly from the above (divide every sample by σ, run the algorithm of Theorem 4, and multiply its output by σ).

As usual in our filtering framework, the algorithm will iteratively look at the top eigenvalue and eigenvector of the sample covariance matrix and return the sample mean if this eigenvalue is small (Algorithm 2). The main difference between this and the filter algorithm for the sub-gaussian case is how we choose the threshold for the filter. Instead of looking for a violation of a concentration inequality, here we will choose a threshold *at random* (with a bias towards higher

thresholds). The reason is that, in this setting, the variance in the direction we look for a filter in needs to be a constant multiple larger – instead of the typical $\tilde{\Omega}(\varepsilon)$ relative for the sub-gaussian case. Therefore, randomly choosing a threshold weighted towards higher thresholds suffices to throw out more corrupted samples than uncorrupted samples *in expectation*. Although it is possible to reject many good samples this way, the algorithm still only rejects a total of $O(\varepsilon)$ samples with high probability.

We would like our good set of samples to have mean close to that of P and bounded variance in all directions. This motivates the following definition:

Definition 4. *We call a set S ε-good for a distribution P with mean μ^P and covariance $\Sigma_P \preceq I$ if the mean μ^S and covariance Σ^S of S satisfy $\|\mu^S - \mu^P\|_2 \leq \sqrt{\varepsilon}$ and $\|\Sigma^S\|_2 \leq 2$.*

However, since we have no assumptions about higher moments, it may be be possible for outliers to affect our sample covariance too much. Fortunately, such outliers have small probability and do not contribute too much to the mean, so we will later reclassify them as errors.

Lemma 5. *Let S be $N = \Theta((d/\varepsilon)\log d)$ samples drawn from P. Then, with probability at least $9/10$, a random $X \in_u S$ satisfies*

(i) $\|\mathbb{E}[X] - \mu^P\|_2 \leq \sqrt{\varepsilon}/3$,

(ii) $\Pr\left[\|X - \mu^P\|_2 \geq 80\sqrt{d/\varepsilon}\right] \leq \varepsilon/160$,

(iii) $\mathbb{E}\left[\|X - \mu^P\|_2 \cdot 1_{\|X-\mu^P\|_2 \geq 80\sqrt{d/\varepsilon}}\right] \leq \sqrt{\varepsilon}/2$, *and*

(iv) $\left\|\mathbb{E}\left[(X-\mu^P)(X-\mu^P)^T \cdot 1_{\|X-\mu^P\|_2 \leq 80\sqrt{d/\varepsilon}}\right]\right\|_2 \leq 3/2$.

Proof. For (i), note that

$$\mathbb{E}_S[\|\mathbb{E}[X] - \mu^P\|_2^2] = \sum_i \mathbb{E}_S[(\mathbb{E}[X]_i - \mu_i^P)^2] \leq d/N \leq \varepsilon/360\,,$$

and so by Markov's inequality, with probability at least $39/40$, we have $\|\mathbb{E}[X] - \mu^P\|_2^2 \leq \varepsilon/9$.

For (ii), similarly to (i), note that

$$\mathbb{E}[\|Y - \mu^P\|_2^2] = \sum_i \mathbb{E}[(Y_i - \mu_i^P)^2] \leq d\,,$$

for $Y \sim P$. By Markov's inequality, $\Pr[\|Y - \mu^P\|_2 \geq 80\sqrt{d/\varepsilon}] \leq \varepsilon/160$ with probability at least $39/40$.

For (iii), note that by an application of the Cauchy-Schwarz inequality

$$\mathbb{E}[|Y - \mu^P| 1_{\|Y-\mu^P\|_2 \geq 80\sqrt{d/\varepsilon}}] \leq \sqrt{\mathbb{E}[(Y-\mu^P)^2]\Pr[\|Y-\mu^P\|_2 \geq 80\sqrt{d/\varepsilon}]} \leq \sqrt{\varepsilon}/80\,.$$

Thus,

$$\mathbb{E}_S[\mathbb{E}[|X - \mu^P| 1_{\|X-\mu^P\|_2 \geq 80\sqrt{d/\varepsilon}}]] \leq \sqrt{\varepsilon}/80\,,$$

and by Markov's inequality, with probability at least $39/40$

$$\mathbb{E}\left[|X - \mu^P| 1_{\|X-\mu^P\|_2 \geq 80\sqrt{d/\varepsilon}}\right] \leq \sqrt{\varepsilon}/2\,.$$

For (iv), we require the following Matrix Chernoff bound:

Lemma 6 (Part of Theorem 5.1.1 of [T$^+$15])**.** *Consider a sequence of $d \times d$ positive semi-definite random matrices X_k with $\|X_k\|_2 \le L$ for all k. Let $\mu^{\max} = \|\sum_k \mathbb{E}[X_k]\|_2$. Then, for $\theta > 0$,*

$$\mathbb{E}\left[\left\|\sum_k X_k\right\|_2\right] \le (e^\theta - 1)\mu^{\max}/\theta + L\log(d)/\theta \ ,$$

and for any $\delta > 0$,

$$\Pr\left[\left\|\sum_k X_k\right\|_2 \ge (1+\delta)\mu^{\max}\right] \le d(e^\delta/(1+\delta)^{1+\delta})^{\mu^{\max}/L} \ .$$

We apply this lemma with $X_k = (x_k - \mu^P)(x_k - \mu^P)^T 1_{\|x_k - \mu^P\|_2 \le 80\sqrt{d/\varepsilon}}$ for $\{x_1, \dots, x_N\} = S$. Note that $\|X_k\|_2 \le (80)^2 d/\varepsilon = L$ and that $\mu^{\max} \le N\|\Sigma_P\|_2 \le N$.

Suppose that $\mu^{\max} \le N/80$. Then, taking $\theta = 1$, we have

$$\mathbb{E}[\left\|\sum_k X_k\right\|_2 \ ] \le (e-1)N/80 + O(d\log(d)/\varepsilon) \ .$$

By Markov's inequality, except with probability $39/40$, we have $\|\sum_k X_k\|_2 \le N + O(d\log(d)/\varepsilon) \le 3N/2$, for N a sufficiently high multiple of $d\log(d)/\varepsilon$.

Suppose that $\mu^{\max} \ge N/80$, then we take $\delta = 1/2$ and obtain

$$\Pr\left[\left\|\sum_k X_k\right\|_2 \ge 3\mu^{\max}2\right] \le d(e^{3/2}/(5/2)^{3/2})^{N\varepsilon/20d} \ .$$

For N a sufficiently high multiple of $d\log(d)/\varepsilon$, we get that $\Pr[\|\sum_k X_k\|_2 \ge 3\mu^{\max}/2] \le 1/40$. Since $\mu^{\max} \le N$, we have with probability at least $39/40$, $\|\sum_k X_k\|_2 \le 3N/2$.

Noting that $\|\sum_k X_k\|_2 /N = \|\mathbb{E}[1_{\|X-\mu^P\|_2 \le 80\sqrt{d/\varepsilon}}(X-\mu^P)(X-\mu^P)^T]\|_2$, we obtain (iv). By a union bound, (i)-(iv) all hold simultaneously with probability at least $9/10$. □

Now we can get a 2ε-corrupted good set from an ε-corrupted set of samples satisfying Lemma 5, by reclassifying outliers as errors:

Lemma 7. *Let $S = R \cup E \setminus L$, where R is a set of $N = \Theta(d\log d/\varepsilon)$ samples drawn from P and E and L are disjoint sets with $|E|, |L| \le \varepsilon$. Then, with probability $9/10$, we can also write $S = G \cup E' \setminus L'$, where $G \subseteq R$ is ε-good, $L' \subseteq L$ and $E' \subseteq E'$ has $|E'| \le 2\varepsilon|S|$.*

Proof. Let $G = \{x \in R : \|x\|_2 \le 80\sqrt{d/\varepsilon}\}$. Since R satisfies (ii) of Lemma 5, $|G| - |R| \le \varepsilon|R|/160 \le \varepsilon|S|$. Thus, $E' = E \cup (R \setminus G)$ has $|E'| \le 3\varepsilon/2$. Note that (iv) of Lemma 5 for R in terms of G is exactly $|G|\|\Sigma^G\|_2/|R| \le 3/2$, and so $\|\Sigma^G\|_2 \le 3|R|/2|G| \le 2$.

It remains to check that $\|\mu^G - \mu^P\|_2 \le \sqrt{\varepsilon}$. But note that (iii) of Lemma 5 is exactly $\mathbb{E}_{X\in uR}[\|X - \mu^P\|_2 1_{X \in R\setminus G}] \le \sqrt{\varepsilon}/2$, and we have

$$\left||G|\mathbb{E}_{X\in uG}[\|X-\mu^P\|_2] - |R|\mathbb{E}_{X\in uR}[\|X-\mu^P\|_2]\right| \le |R|\sqrt{\varepsilon}/2 \ ,$$

and since by (i), $\mathbb{E}_{X\in uR}[\|X-\mu^P\|_2] \le \sqrt{\varepsilon}/3$, it follows that $\mathbb{E}_{X\in uG'}[\|X-\mu^P\|_2] \le \sqrt{\varepsilon}$. □

An iteration of FilterUnder2ndMoment may throw out more samples from G than corrupted samples. However, in expectation, we throw out many more corrupted samples than from the good set:

Algorithm 2 Filter under second moment assumptions

function FilterUnder2ndMoment(S)
  Compute μ^S, Σ^S, the mean and covariance matrix of S.
  Find the eigenvector v^* with highest eigenvalue λ^* of Σ^S.
  if $\lambda^* \leq 9$ **then**
    return μ^S
  else
    Draw Z from the distribution on $[0,1]$ with probability density function $2x$.
    Let $T = Z \max\{|v^* \cdot x - \mu^S| : x \in S\}$.
    Return the set $S' = \{x \in S : |v^* \cdot (X - \mu^S)| < T\}$.
  end if
end function

Proposition 2. *If we run* F*ilterUnder2ndMoment on a set $S = G \cup E \setminus L$ for some ε-good set G and disjoint E, L with $|E| \leq 2\varepsilon|S|, |L| \leq 9\varepsilon|S|$, then either it returns μ^S with $\|\mu^S - \mu^P\|_2 \leq O(\sqrt{\varepsilon})$, or else it returns a set $S' \subset S$ with $S' = G \cup E' \setminus L'$ for disjoint E' and L'. In the latter case we have $\mathbb{E}_Z[|E'| + 2|L'|] \leq |E| + 2|L|$.*

For $D \in \{G, E, L, S\}$, let μ^D be the mean of D and M_D be the matrix $\mathbb{E}_{X \in_u D}[(X - \mu^S)(X - \mu^S)^T]$.

Lemma 8. *If G is an ε-good set with $x \leq 40\sqrt{d/\varepsilon}$ for $x \in S \cup G$, then $\|M_G\|_2 \leq 2\|\mu^G - \mu^S\|_2^2 + 2$.*

Proof. For any unit vector v, we have

$$\begin{aligned} v^T M_G v &= \mathbb{E}_{X \in_u G}[(v \cdot (X - \mu^S))^2] \\ &= \mathbb{E}_{X \in_u G}[(v \cdot (X - \mu^G) + v \cdot (\mu^P - \mu^G))^2] \\ &= v^T \Sigma^G v + (v \cdot (\mu^G - \mu^S))^2 \\ &\leq 2 + 2\|\mu^G - \mu^S\|_2^2 \, . \end{aligned}$$

□

Lemma 9. *We have that $|L|\|M_L\|_2 \leq 2|G|(1 + \|\mu^G - \mu^S\|_2^2)$.*

Proof. Since $L \subseteq G$, for any unit vector v, we have

$$\begin{aligned} |L| v^T M_L v &= |L| \mathbb{E}_{X \in_u L}[(v \cdot (X - \mu^S))^2] \\ &\leq |G| \mathbb{E}_{X \in_u G}[(v \cdot (X - \mu^S))^2] \\ &\leq 2|G|(1 + \|\mu^G - \mu^S\|_2^2) \, . \end{aligned}$$

□

Lemma 10. $\|\mu^G - \mu^S\|_2 \leq \sqrt{2\varepsilon\|M_S\|_2} + 12\sqrt{\varepsilon}$.

Proof. We have that $|E|M_E \leq |S|M_S + |L|M_L$ and so

$$|E|\|M_E\|_2 \leq |S|\|M_S\|_2 + 2|G|(1 + \|\mu^G - \mu^S\|_2^2) \, .$$

By Cauchy Schwarz, we have that $\|M_E\|_2 \geq \|\mu^E - \mu^S\|_2^2$, and so

$$\sqrt{|E|}\|\mu^E - \mu^S\|_2 \leq \sqrt{|S|\|M_S\|_2 + 2|G|(1 + \|\mu^G - \mu^S\|_2^2)} \, .$$

By Cauchy-Schwarz and Lemma 9, we have that

$$\sqrt{|L|}\|\mu^L - \mu^S\|_2 \leq \sqrt{|L|\|M_L\|_2} \leq \sqrt{2|G|(1 + \|\mu^G - \mu^S\|_2^2)} \, .$$

Since $|S|\mu^S = |G|\mu^G + |E|\mu^E - |L|\mu^L$ and $|S| = |G| + |E| - |L|$, we get

$$|G|(\mu^G - \mu^S) = |E|(\mu^E - \mu^S) - |L|(\mu^E - \mu^S) \, .$$

Substituting into this, we obtain

$$|G|\|\mu^G - \mu^S\|_2 \leq \sqrt{|E||S|\|M_S\|_2 + 2|E||G|(1 + \|\mu^G - \mu^S\|_2^2)} + \sqrt{2|L||G|(1 + \|\mu^G - \mu^S\|_2^2)} \, .$$

Since for $x, y > 0$, $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, we have

$$|G|\|\mu^G - \mu^S\|_2 \leq \sqrt{|E||S|\|M_S\|_2} + (\sqrt{2|E||G|} + \sqrt{2|L||G|})(1 + \|\mu^G - \mu^S\|_2) \, .$$

Since $||G| - |S|| \leq \varepsilon|S|$ and $|E| \leq 2\varepsilon|S|, |L| \leq 9\varepsilon|S|$, we have

$$\|\mu^G - \mu^S\|_2 \leq \sqrt{2\varepsilon\|M_S\|_2} + (6\sqrt{\varepsilon})(1 + \|\mu^G - \mu^S\|_2) \, .$$

Moving the $\|\mu^G - \mu^S\|_2$ terms to the LHS, using $6\sqrt{\varepsilon} \leq 1/2$, gives

$$\|\mu^G - \mu^S\|_2 \leq \sqrt{2\varepsilon\|M_S\|_2} + 12\sqrt{\varepsilon} \, .$$

□

Since $\lambda^* = \|M_S\|_2$, the correctness if we return the empirical mean is immediate.

Corollary 3. *If $\lambda^* \leq 9$, we have that $\|\mu^G - \mu^S\|_2 = O(\sqrt{\varepsilon})$.*

From now on, we assume $\lambda^* > 9$. In this case we have $\|\mu^G - \mu^S\|_2^2 \leq O(\varepsilon\lambda^*)$. Using Lemma 8, we have

$$\|M_G\|_2 \leq 2 + O(\varepsilon\lambda^*) \leq 2 + \lambda^*/5$$

for sufficiently small ε. Thus, we have that

$$v^{*T} M_S v^* \geq 4 v^{*T} M_G v^* \, . \tag{7}$$

Now we can show that in expectation, we throw out many more corrupted points from E than from $G \setminus L$:

Lemma 11. *Let $S' = G \cup E' \setminus L'$ for disjoint E', L' be the set of samples returned by the iteration. Then we have $\mathbb{E}_Z[|E'| + 2|L'|] \leq |E| + 2|L|$.*

Proof. Let $a = \max_{x \in S} |v^* \cdot x - \mu^S|$. Firstly, we look at the expected number of samples we reject:

$$\begin{aligned}
\mathbb{E}_Z[|S'|] - |S| &= \mathbb{E}_Z\left[|S| \Pr_{X \in_u S}[|X - \mu^S| \geq aZ]\right] \\
&= |S| \int_0^1 \Pr_{X \in_u S}\left[|v^* \cdot (X - \mu^S)| \geq ax\right] 2x dx \\
&= |S| \int_0^a \Pr_{X \in_u S}\left[|v^* \cdot (X - \mu^S)| \geq T\right] (2T/a) dT \\
&= |S| \mathbb{E}_{X \in_u S}\left[(v^* \cdot (X - \mu^S))^2\right]/a \\
&= (|S|/a) \cdot v^{*T} M_S v^* \, .
\end{aligned}$$

Next, we look at the expected number of false positive samples we reject, i.e., those in $L' \setminus L$.

$$\begin{aligned}
\mathbb{E}_Z[|L'|] - |L| &= \mathbb{E}_Z\left[(|G| - |L|) \Pr_{X \in_u G \setminus L}\left[|X - \mu^S| \geq T\right]\right] \\
&\leq \mathbb{E}_Z\left[|G| \Pr_{X \in_u G}[|v^* \cdot (X - \mu^S)| \geq aZ]\right] \\
&= |G| \int_0^1 \Pr_{X \in_u G}[|v^* \cdot (X - \mu^S)| \geq ax] 2x \, dx \\
&= |G| \int_0^a \Pr_{X \in_u G}[|v^* \cdot (X - \mu^S)| \geq T](2T/a) \, dT \\
&\leq |G| \int_0^\infty \Pr_{X \in_u G}[|v^* \cdot (X - \mu^S)| \geq T](2T/a) \, dT \\
&= |G| \mathbb{E}_{X \in_u G}\left[(v^* \cdot (X - \mu^S))^2\right] / a \\
&= (|G|/a) \cdot v^{*T} M_G v^* \, .
\end{aligned}$$

Using (7), we have $|S| v^{*T} M_S v^* \geq 4|G| v^{*T} M_G v^*$ and so $\mathbb{E}_Z[S'] - S \geq 3(\mathbb{E}_Z[L'] - L)$. Now consider that $|S'| = |G| + |E'| - |L'| = |S| - |E| + |E'| + |L| - |L'|$, and thus $|S'| - |S| = |E| - |E'| + |L'| - |L|$. This yields that $|E| - \mathbb{E}_Z[|E'|] \geq 2(\mathbb{E}_Z[L'] - L)$, which can be rearranged to $\mathbb{E}_Z[|E'| + 2|L'|] \leq |E| + 2|L|$. □

Proof of Proposition 2. If $\lambda^* \leq 9$, then we return the mean in Step 5, and by Corollary 3, $\|\mu^S - \mu^P\|_2 \leq O(\sqrt{\varepsilon})$.

If $\lambda^* > 9$, then we return S'. Since at least one element of S has $|v^* \cdot X| = \max_{x \in S} |v^* \cdot X|$, whatever value of Z is drawn, we still remove at least one element, and so have $S' \subset S$. By Lemma 11, we have $\mathbb{E}_Z[|E'| + 2|L'|] \leq |E| + 2|L|$. □

Proof of Theorem 4. Our input is a set S of $N = \Theta((d/\varepsilon) \log d)$ ε-corrupted samples so that with probability 9/10, S is a 2ε-corrupted set of ε-good samples for P by Lemmas 5 and 7. We have a set $S = G \cup E' \setminus L$, where G' is an ε-good set, $|E| \leq 2\varepsilon$, and $|L| \leq \varepsilon$. Then, we iteratively apply FilterUnder2ndMoment until it outputs an approximation to the mean. Since each iteration removes a sample, this must happen within N iterations. The algorithm takes at most $\mathrm{poly}(N, d) = \mathrm{poly}(d, 1/\varepsilon)$ time.

As long as we can show that the conditions of Proposition 2 hold in each iteration, it ensures that $\|\mu^S - \mu^P\|_2 \leq O(\sqrt{\varepsilon})$. However, the condition that $|L| \leq 9\varepsilon|S|$ need not hold in general. Although in expectation we reject many more samples in E than G, it is possible that we are unlucky and reject many samples in G, which could make L large in the next iteration. Thus, we need a bound on the probability that we ever have $|L| > 9\varepsilon$.

We analyze the following procedure: We iteratively run FilterUnder2ndMoment starting with a set $S_i \cup E_i \setminus L_i$ of samples with $S_0 = S$ and producing a set $S_{i+1} = G \cup E_{i+1} \setminus L_{i+1}$. We stop if we output an approximation to the mean or if $|L_{i+1}| \geq 13\varepsilon|S|$. Since we do now always satisfy the conditions of Proposition 2, this gives that $\mathbb{E}_Z[|E_{i+1}| + |L_{i+1}|] = |E_i| + 2|L_i|$. This expectation is conditioned on the state of the algorithm after previous iterations, which is determined by S_i. Thus, if we consider the random variables $X_i = |E_i| + 2|L_i|$, then we have $\mathbb{E}[X_{i+1}|S_i] \leq X_i$, i.e., the sequence X_i is a sub-martingale with respect to X_i. Using the convention that $S_{i+1} = S_i$, if we stop in less than i iterations, and recalling that we always stop in N iterations, the algorithm fails if and only if $|L_N| > 9\varepsilon|S|$. By a simple induction or standard results on sub-martingales, we have $\mathbb{E}[X_N] \leq X_0$. Now $X_0 = |E_0| + 2|L_0| \leq 3\varepsilon|S|$. Thus, $\mathbb{E}[X_N] \leq 3\varepsilon|S|$. By Markov's inequality, except with probability 1/6, we have $X_N \leq 18\varepsilon|S|$. In this case, $|L_N| \leq X_N/2 \leq 9\varepsilon|S|$. Therefore, the probability that we ever have $|L_i| > 9\varepsilon$ is at most 1/6.

By a union bound, the probability that the uncorrupted samples satisfy Lemma 5 and Proposition 2 applies to every iteration is at least $9/10 - 1/6 \geq 2/3$. Thus, with at least $2/3$ probability, the algorithm outputs a vector $\widehat{\mu}$ with $\|\widehat{\mu} - \mu^P\|_2 \leq O(\sqrt{\varepsilon})$. □

1.3 Robust Covariance Estimation

In this subsection, we give a near sample-optimal efficient robust estimator for the covariance of a zero-mean Gaussian density, thus proving Theorem 3.3. Our algorithm is essentially identical to the filtering algorithm given in Section 8.2 of [DKK+16]. As in Section 1.1 the only difference is a weaker definition of the "good set of samples" (Definition 5) and a concentration argument (Lemma 3) showing that a random set of uncorrupted samples of the appropriate size is good with high probability. Given these, the analysis of this subsection follows straightforwardly from the analysis in Section 8.2 of [DKK+16] by plugging in the modified parameters.

The algorithm Filter-Gaussian-Unknown-Covariance to robustly estimate the covariance of a mean 0 Gaussian in [DKK+16] is as follows:

Algorithm 3 Filter algorithm for a Gaussian with unknown covariance matrix.

procedure Filter-Gaussian-Unknown-Covariance(S', ε, τ)
input: A multiset S' such that there exists an (ε, τ)-good set S with $\Delta(S, S') \leq 2\varepsilon$
output: Either a set S'' with $\Delta(S, S'') < \Delta(S, S')$ or the parameters of a Gaussian G' with $d_{TV}(G, G') = O(\epsilon \log(1/\epsilon))$.
Let $C > 0$ be a sufficiently large universal constant.
Let Σ' be the matrix $\mathbb{E}_{X \in_u S'}[XX^T]$ and let G' be the mean 0 Gaussian with covariance matrix Σ'.
if there is any $x \in S'$ so that $x^T(\Sigma')^{-1}x \geq Cd\log(|S'|/\tau)$ **then**
return $S'' = S' - \{x : x^T(\Sigma')^{-1}x \geq Cd\log(|S'|/\tau)\}$.
end if
Compute an approximate eigendecomposition of Σ' and use it to compute $\Sigma'^{-1/2}$
Let $x_{(1)}, \ldots, x_{(|S'|)}$ be the elements of S'.
For $i = 1, \ldots, |S'|$, let $y_{(i)} = \Sigma'^{-1/2}x_{(i)}$ and $z_{(i)} = y_{(i)}^{\otimes 2}$.
Let $T_{S'} = -I^\flat I^{\flat T} + (1/|S'|)\sum_{i=1}^{|S'|} z_{(i)}z_{(i)}^T$.
Approximate the top eigenvalue λ^* and corresponding unit eigenvector v^* of $T_{S'}$..
Let $p^*(x) = \frac{1}{\sqrt{2}}((\Sigma'^{-1/2}x)^T v^{*\sharp}(\Sigma'^{-1/2}x) - \mathrm{tr}(v^{*\sharp}))$
if $\lambda^* \leq (1 + C\epsilon\log^2(1/\epsilon))Q_{G'}(p^*)$ **then**
return G'
end if
Let μ be the median value of $p^*(X)$ over $X \in S'$.
Find a $T \geq C'$ so that

$$\Pr_{X \in_u S'}(|p^*(X) - \mu| \geq T + 4/3) \geq \mathrm{Tail}(T, d, \varepsilon, \tau)$$

return $S'' = \{X \in S' : |p^*(X) - \mu| < T\}$.
end procedure

In [DKK+16], we take $\mathrm{Tail}(T, d, \varepsilon, \tau) = 12\exp(-T) + 3\epsilon/(d\log(N/\tau))^2$, where $N = \Theta((d\log(d/\varepsilon\tau))^6/\varepsilon^2)$ is the number of samples we took there.

To get a near sample-optimal algorithms, we will need a weaker definition of a good set. To use this, we will need to weaken the tail bound in the algorithm to $\mathrm{Tail}(T, d, \varepsilon, \tau) = \varepsilon/(T^2 \log^2(T))$, when $T \geq 10\log(1/\varepsilon)$. For $T \leq 10\log(1/\varepsilon)$, we take $\mathrm{Tail}(T, d, \varepsilon, \tau) = 1$ so that we always choose $T \geq 10\log(1/\varepsilon)$. It is easy to show that the integrals of this tail bound used in the proofs of Lemma 8.19 and Claim 8.22 of [DKK+16] have similar bounds. Thus, our analysis here will sketch that these tail bounds hold for a set of $\Omega(d^2 \log^5(d/\varepsilon\tau)/\varepsilon^2)$ samples from the Guassian.

Firstly, we state the new, weaker, definition of a good set:

Definition 5. *Let G be a Gaussian in R^d with mean 0 and covariance Σ. Let $\epsilon > 0$ be sufficiently small. We say that a multiset S of points in R^d is ε-good with respect to G if the following hold:*

1. *For all $x \in S$, $x^T \Sigma^{-1} x < d + O(\sqrt{d}\log(d/\varepsilon))$.*
2. *We have that $\|\Sigma^{-1/2}\mathrm{Cov}(S)\Sigma^{-1/2} - I\|_F = O(\varepsilon)$.*
3. *For all even degree-2 polynomials p, we have that $\mathrm{Var}(p(S)) = \mathrm{Var}(p(G))(1 + O(\varepsilon))$.*
4. *For p an even degree-2 polynomial with $\mathbb{E}[p(G)] = 0$ and $\mathrm{Var}(p(G)) = 1$, and for any $T > 10\log(1/\varepsilon)$ we have that*
$$\Pr_{x \in_u S}(|p(x)| > T) \leq \varepsilon/(T^2 \log^2(T)).$$

It is easy to see that the algorithm and analysis of [DKK+16] can be pushed through using the above weaker definition. That is, if S is a good set, then G can be recovered to $\tilde{O}(\varepsilon)$ error from an ε-corrupted version of S. Our main task will be to show that random sets of the appropriate size are good with high probability.

Proposition 3. *Let N be a sufficiently large constant multiple of $d^2 \log^5(d/\varepsilon)/\varepsilon^2$. Then a set S of N independent samples from G is ε-good with respect to G with high probability.*

Proof. First, note that it suffices to prove this when $G = N(0, I)$.

Condition 1 follows by standard concentration bounds on $\|x\|_2^2$.

Condition 2 follows by estimating the entry-wise error between $\mathrm{Cov}(S)$ and I.

Condition 3 is slightly more involved. Let $\{p_i\}$ be an orthonormal basis for the set of even, degree-2, mean-0 polynomials with respect to G. Define the matrix $M_{i,j} = \mathbb{E}_{x \in_u S}[p_i(x)p_j(x)] - \delta_{i,j}$. This condition is equivalent to $\|M\|_2 = O(\varepsilon)$. Thus, it suffices to show that for every v with $\|v\|_2 = 1$ that $v^T M v = O(\varepsilon)$. It actually suffices to consider a cover of such v's. Note that this cover will be of size $2^{O(d^2)}$. For each v, let $p_v = \sum_i v_i p_i$. We need to show that $\mathrm{Var}(p_v(S)) = 1 + O(\varepsilon)$. We can show this happens with probability $1 - 2^{-\Omega(d^2)}$, and thus it holds for all v in our cover by a union bound.

Condition 4 is substantially the most difficult of these conditions to prove. Naively, we would want to find a cover of all possible p and all possible T, and bound the probability that the desired condition fails. Unfortunately, the best a priori bound on $\Pr(|p(G)| > T)$ are on the order of $\exp(-T)$. As our cover would need to be of size 2^{d^2} or so, to make this work with $T = d$, we would require on the order of d^3 samples in order to make this argument work.

However, we will note that this argument is sufficient to cover the case of $T < 10\log(1/\varepsilon)\log^2(d/\varepsilon)$.

Fortunately, most such polynomials p satisfy much better tail bounds. Note that any even, mean zero polynomial p can be written in the form $p(x) = x^T A x - \mathrm{tr}(A)$ for some matrix A. We call A the associated matrix to p. We note by the Hanson-Wright inequality that $\Pr(|p(G)| > T) = \exp(-\Omega(\min((T/\|A\|_F)^2, T/\|A\|_2)))$. Therefore, the tail bounds above are only as bad as described when A has a single large eigenvalue. To take advantage of this, we will need to break p into parts based on the size of its eigenvalues. We begin with a definition:

Definition 6. *Let $\mathcal{P}_k$ be the set of even, mean-0, degree-2 polynomials, so that the associated matrix A satisfies:*

1. $\text{rank}(A) \leq k$
2. $\|A\|_2 \leq 1/\sqrt{k}$.

Note that for $p \in \mathcal{P}_k$ that $|p(x)| \leq |x|^2/\sqrt{k} + \sqrt{k}$.

Importantly, any polynomial can be written in terms of these sets.

Lemma 12. *Let p be an even, degree-2 polynomial with $\mathbb{E}[p(G)] = 0, \text{Var}(p(G)) = 1$. Then if $t = \lfloor \log_2(d) \rfloor$, it is possible to write $p = 2(p_1 + p_2 + \ldots + p_{2^t} + p_d)$ where $p_k \in \mathcal{P}_k$.*

Proof. Let A be the associated matrix to p. Note that $\|A\|_F = \text{Var}p = 1$. Let A_k be the matrix corresponding to the top k eigenvalues of A. We now let p_1 be the polynomial associated to $A_1/2$, p_2 be associated to $(A_2 - A_1)/2$, p_4 be associated to $(A_4 - A_2)/2$, and so on. It is clear that $p = 2(p_1 + p_2 + \ldots + p_{2^t} + p_d)$. It is also clear that the matrix associated to p_k has rank at most k. If the matrix associated to p_k had an eigenvalue more than $1/\sqrt{k}$, it would need to be the case that the $k/2^{nd}$ largest eigenvalue of A had size at least $2/\sqrt{k}$. This is impossible since the sum of the squares of the eigenvalues of A is at most 1.

This completes our proof. □

We will also need covers of each of these sets $\mathcal{P}_k$.

Lemma 13. *For each k, there exists a set $\mathcal{C}_k \subset \mathcal{P}_k$ so that*

1. *For each $p \in \mathcal{P}_k$ there exists a $q \in \mathcal{C}_k$ so that $\|p(G) - q(G)\|_2 \leq (\varepsilon/d)^2$.*
2. $|\mathcal{C}_k| = 2^{O(dk \log(d/\varepsilon))}$.

Proof. We note that any such p is associated to a matrix A of the form $A = \sum_{i=1}^k \lambda_i v_i v_i^T$, for $\lambda_i \in [0, 1/\sqrt{k}]$ and v_i orthonormal. It suffices to let q correspond to the matrix $A' = \sum_{i=1}^k \mu_i w_i w_i^T$ for with $|\lambda_i - \mu_i| < (\varepsilon/d)^3$ and $|v_i - w_i| < (\varepsilon/d)^3$ for all i. It is easy to let μ_i and w_i range over covers of the interval and the sphere with appropriate errors. This gives a set of possible q's of size $2^{O(dk \log(d/\varepsilon))}$ as desired. Unfortunately, some of these q will not be in $\mathcal{P}_k$ as they will have eigenvalues that are too large. However, this is easily fixed by replacing each such q by the closest element of $\mathcal{P}_k$. This completes our proof. □

We next will show that these covers are sufficient to express any polynomial.

Lemma 14. *Let p be an even degree-2 polynomial with $\mathbb{E}[p(G)] = 0$ and $\text{Var}(p(G)) = 1$. It is possible to write p as a sum of $O(\log(d))$ elements of some $\mathcal{C}_k$ plus another polynomial of L^2 norm at most ε/d.*

Proof. Combining the above two lemmas we have that any such p can be written as

$$p = (q_1 + p_1) + (q_2 + p_2) + \ldots (q_{2^t} + p_{2^t}) + (q_d + p_d) = q_1 + q_2 + \ldots + q^{2^t} + q^d + p' \,,$$

where q_k above is in $\mathcal{C}_k$ and $\|p_k(G)\|_2 < (\varepsilon/d)^2$. Thus, $p' = p_1 + p_2 + \ldots + p_{2^t} + p_d$ has $\|p'(G)\|_2 \leq (\varepsilon/d)$. This completes the proof. □

The key observation now is that if $|p(x)| \geq T$ for $\|x\|_2 \leq \sqrt{d/\varepsilon}$, then writing $p = q_1 + q_2 + q_4 + \ldots + q_d + p'$ as above, it must be the case that $|q_k(x)| > (T-1)/(2\log(d))$ for some k. Therefore, to prove our main result, it suffices to show that, with high probability over the choice of S, for any $T \geq 10\log(1/\varepsilon)\log^2(d/\varepsilon)$ and any $q \in \mathcal{C}_k$ for some k, that $\Pr_{x\in_u S}(|q(x)| > T/(2\log(d))) < \varepsilon/(2T^2\log^2(T)\log(d))$. Equivalently, it suffices to show that for $T \geq 10\log(1/\varepsilon)\log(d/\varepsilon)$ it holds $\Pr_{x\in_u S}(|q(x)| > T/(2\log(d))) < \varepsilon/(2T^2\log^2(T)\log^2(d))$. Note that this holds automatically for $T > (d/\varepsilon)$, as $p(x)$ cannot possibly be that large for $\|x\|_2 \leq \sqrt{d/\varepsilon}$. Furthermore, note that losing a constant factor in the probability, it suffices to show this only for T a power of 2.

Therefore, it suffices to show for every $k \leq d$, every $q \in \mathcal{C}_k$ and every $d/\sqrt{k\varepsilon} \gg T \gg \log(1/\varepsilon)\log(d/\varepsilon)$ that with probability at least $1 - 2^{-\Omega(dk\log(d/\varepsilon))}$ over the choice of S we have that $\Pr_{x\in_u S}(|q(x)| > T) \ll \varepsilon/(T^2\log^4(d/\varepsilon))$. However, by the Hanson-Wright inequality, we have that

$$\Pr(|q(G)| > T) = \exp(-\Omega(\min(T^2, T\sqrt{k}))) < (\varepsilon/(T^2\log^4(d/\varepsilon)))^2 .$$

Therefore, by Chernoff bounds, the probability that more than a $\varepsilon/(T^2\log^4(d/\varepsilon))$-fraction of the elements of S satisfy this property is at most

$$\begin{aligned}\exp(-\Omega(\min(T^2, T\sqrt{k}))|S|\varepsilon/(T^2\log^4(d/\varepsilon))) &= \exp(-\Omega(|S|\varepsilon/(\log^4(d/\varepsilon))\min(1, \sqrt{k}/T))) \\ &\leq \exp(-\Omega(|S|\varepsilon^2/(\log^4(d/\varepsilon))k/d)) \\ &\leq \exp(-\Omega(dk\log(d/\varepsilon))) ,\end{aligned}$$

as desired.

This completes our proof. □

2 Omitted Details from Section 5

2.1 Full description of the distributions for experiments

Here we formally describe the distributions we used in our experiments. In all settings, our goal was to find noise distributions so that noise points were not "obvious" outliers, in the sense that there is no obvious pointwise pruning process which could throw away the noise points, which still gave the algorithms we tested the most difficulty. We again remark that while other algorithms had varying performances depending on the noise distribution, it seemed that the performance of ours was more or less unaffected by it.

Distribution for the synthetic mean experiment Our uncorrupted points were generated by $\mathcal{N}(\mu, I)$, where μ is the all-ones vector. Our noise distribution is given as

$$N = \frac{1}{2}\Pi_1 + \frac{1}{2}\Pi_2 ,$$

where Π_1 is the product distribution over the hypercube where every coordinate is 0 or 1 with probability 1/2, and Π_2 is a product distribution where the first coordinate is ether 0 or 12 with equal probability, the second coordinate is -2 or 0 with equal probability, and all remaining coordinates are zero.

Distribution for the synthetic covariance experiment For the isotropic synthetic covariance experiment, our uncorrupted points were generated by $\mathcal{N}(0, I)$, and the noise points were all zeros. For the skewed synthetic covariance experiment, our uncorrupted points were generated by $\mathcal{N}(0, I+$

$10e_1e_1^T$), where e_1 is the first unit vector, and our noise points were generated as follows: we took a fixed random rotation of points of the form $Y_i \sim \Pi$, where Π is a product distribution whose first $d/2$ coordinates are ± 0.5 with probability $1/2$, and whose next $d/2 - 1$ coordinates are each $0.8 \times A_i$, where for each coordinate i, A_i is an independent random integer between -2 and 2, and whose last coordinate is a uniformly random integer between $[-10, 10]$.

Setup for the semi-synthetic geographic experiment We took the 20 dimensional data from [NJB+08], which was diagonalized, and randomly rotated it. This was to simulate the higher dimensional case, since the singular vectors that [NJB+08] obtained did not seem to be sparse or analytically sparse. Our noise was distributed as Π, where Π is a product distribution whose first $d/2$ coordinates are each uniformly random integers between 0 and 2 and whose last $d/2$ coordinates are each uniformly randomly either 2 or 3, all scaled by a factor of $1/24$.

2.2 Comparison with other robust PCA methods on semi-synthetic data

In addition to comparing our results with simple pruning techniques, as we did in Figure 3 in the main text, we also compared our algorithm with implementations of other robust PCA techniques from the literature with accessible implementations. In particular, we compared our technique with RANSAC-based techniques, `LRVCov`, two SDPs ([CLMW11, XCS10]) for variants of robust PCA, and an algorithm proposed by [CLMW11] to speed up their SDP based on alternating descent. For the SDPs, since black box methods were too slow to run on the full data set (as [CLMW11] mentions, black-box solvers for the SDPs are impractical above perhaps 100 data points), we subsample the data, and run the SDP on the subsampled data. For each of these methods, we ran the algorithm on the true data points plus noise, where the noise was generated as described above. We then take the estimate of the covariance it outputs, and project the data points onto the top two singular values of this matrix, and plot the results in Figure 1.

Similar results occurred for most noise patterns we tried. We found that only our algorithm and `LRVCov` were able to reasonably reconstruct Europe, in the presence of this noise. It is hard to judge qualitatively which of the two maps generated is preferable, but it seems that ours stretches the picture somewhat less than `LRVCov`.

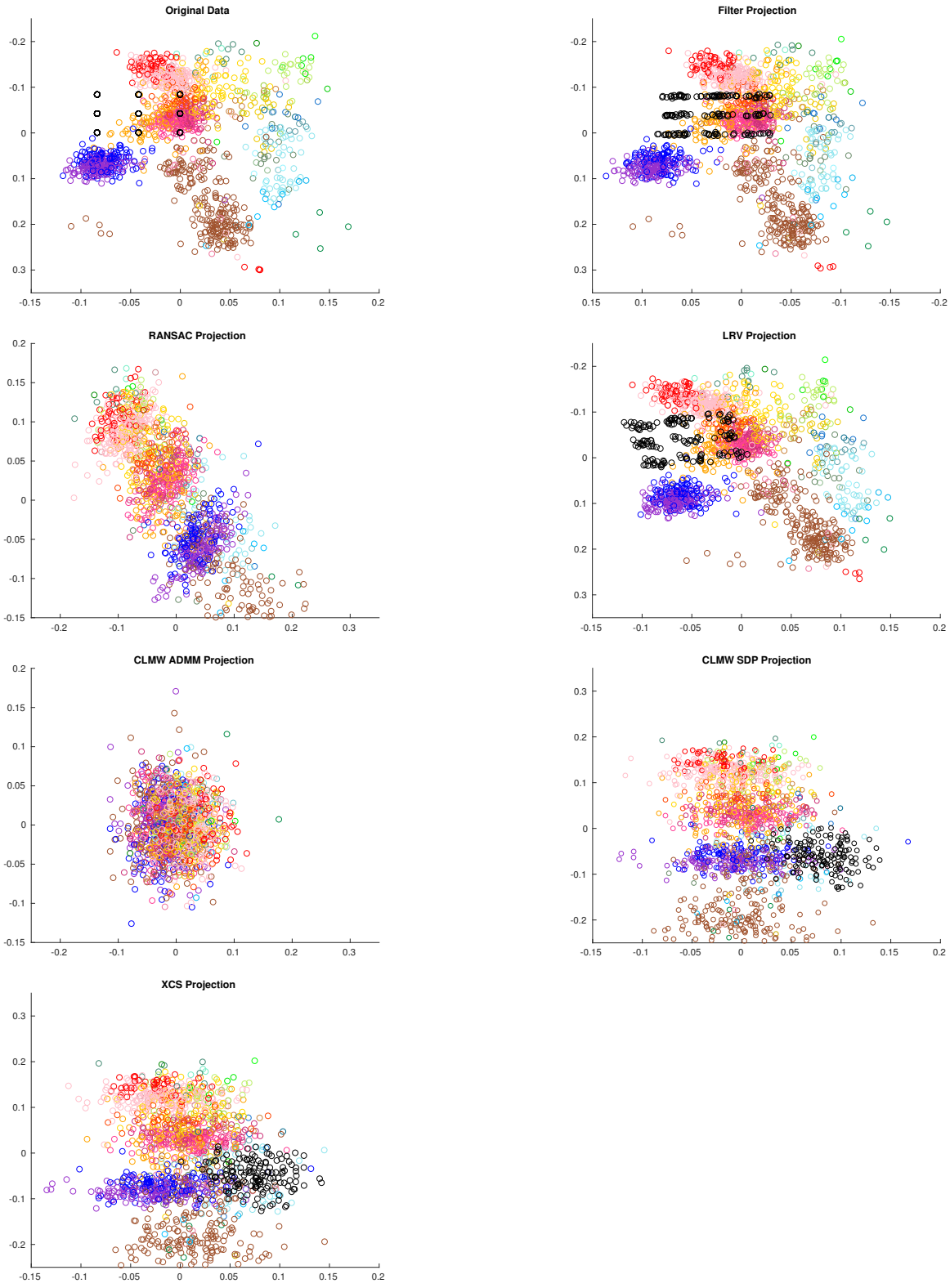

Figure 1: Comparison with other robust methods on the Europe semi-synthetic data. From left to right, top to bottom: the original projection without noise, what our algorithm recovers, RANSAC, `LRVCov`, the ADMM method proposed by [CLMW11], the SDP proposed by [XCS10] with subsampling, and the SDP proposed by [CLMW11] with subsampling.

References

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[DKK+16] 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*, 2016. Full version available at https://arxiv.org/pdf/1604.06443.pdf.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

[NJB+08] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[T+15] J. A. Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[Ver10] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2010.

[XCS10] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.