8. Supplementary Material

We now provide the proofs for all results stated in the main body of our work. We also introduce the connection between recovery for replacements and sparse recovery. Finally, we provide additional details and figures from our experimental evaluation that were omitted due to space constraints.

8.1. Proof of Proposition 1

Proof. Suppose that $\mathcal{A}_3$ affects at least one entry in a subset S of all samples. As at least one coordinate per sample is corrupted, S must be at most an α-fraction of all samples; since $\alpha \leq \epsilon/n$ the sample-level adversary can corrupt the entirety of every sample partially corrupted by the coordinate-level adversary, and thus, it is a stronger adversary given this condition. The proof for $\mathcal{A}_2$ is similar. □

8.2. Proof of Proposition 2

Proof. If $\alpha,\ \rho \geq \epsilon$, similarly to the proof of Proposition 1, $\mathcal{A}_2$ and $\mathcal{A}_3$ can simulate $\mathcal{A}_1$ by placing all its corruptions on the ϵN coordinates corrupted by $\mathcal{A}_1$. If $\alpha \geq \rho$, $\mathcal{A}_3$ can simulate $\mathcal{A}_2$ by corrupting the coordinates corrupted by $\mathcal{A}_2$ since $\mathcal{A}_2$ can never corrupt more than ρ-fraction of coordinates in expectation. On the other hand, if $\alpha \leq \rho/n$, $\mathcal{A}_2$ can corrupt whatever coordinates $\mathcal{A}_3$ decides to corrupt since $\mathcal{A}_3$ cannot corrupt more than αn-fraction of one coordinate. Thus, the three statements hold. □

8.3. Proof of Theorem 1

Proof. We first show that when $\alpha > 2\alpha'$, $\mathcal{A}_3^\alpha$ has a way to make corruptions so that with probability at least $1 - e^{-\Omega(\alpha^2 N)}$ it is indistinguishable whether the N samples come from D_1 or D_2. From the definition of d^1_{ENTRY}, if we take the coupling γ that achieves the infimum, changing α' fraction of the entries per sample on average will make D_1 indistinguishable from D_2. Therefore, if the adversary corrupts the entries of the N samples according to the coupling γ, by Hoeffding's inequality, the probability that more than $2\alpha'$ fraction of the entries need to be changed to make it impossible to tell whether the samples come from D_1 or D_2 is less than $e^{-\Omega(\alpha^2 N)}$.

Then we show that when $\alpha < \alpha'/4$, no matter how $\mathcal{A}_3^\alpha$ makes corruptions, with probability at least $1 - e^{-\Omega(\alpha^2 N)}$, we can tell that the N samples come from D_1. Since $d^1_{\text{ENTRY}}(D_1, D_2) = \alpha'$, by Monge-Kantorovich duality theorem (see e.g. Theorem 5.10 of (Villani, 2009)), there exists a function $u : (\mathbb{R} \cup \{\perp\})^n \to [0, 1]$, where $\perp$ denotes a missing entry, such that $u(x) - u(y) \leq \frac{1}{n}\|x - y\|_0$ and $\mathbb{E}_{D_1}[u(x)] - \mathbb{E}_{D_2}[u(x)] = \alpha'$. This is because the optimal coupling of D_1, D_2 for d^1_{ENTRY} is represents the optimal to the primal Kantorovich problem where $c(x, y) = \frac{1}{n}\|I(x, y)\|_1$, while u represents the optimal to the dual problem. We use u to distinguish whether the corrupted samples come from D_1 or D_2 by checking whether the expectation of u according to the empirical distribution $\hat{D}$ that we observe is closer to the expectation corresponding to D_1 or D_2. By Hoeffding's inequality, the empirical distribution $\hat{D}_1$ of the N samples before corruption satisfies $|\mathbb{E}_{D_1}[u(x)] - \mathbb{E}_{\hat{D}_1}[u(x)]| \leq \alpha'/4$ with probability at least $1 - e^{-\Omega(\alpha^2 N)}$. After corruption, we have that $|\mathbb{E}_{\hat{D}}[u(x)] - \mathbb{E}_{\hat{D}_1}[u(x)]| \leq \alpha$ by the bound on the number of corrupted entries and the Lipschitz property of u. Thus, with probability at least $1 - e^{-\Omega(\alpha^2 N)}$, $|\mathbb{E}_{D_1}[u(x)] - \mathbb{E}_{\hat{D}}[u(x)]| \leq \alpha'/4 + \alpha < \alpha'/2$ while $|\mathbb{E}_{D_2}[u(x)] - \mathbb{E}_{\hat{D}}[u(x)]| > \alpha'/2$, which allows us to distinguish between D_1 and D_2.

In the case of the value-fraction adversary $\mathcal{A}_2^\rho$ that can corrupt ρ-fraction of values in each coordinate, d^∞_{ENTRY} can be bound similarly in $\Theta(\rho)$ by applying Hoeffding's inequality and Kantorovich duality theorem for each coordinate such that $u_i(x) - u_i(y) \leq \|x_i - y_i\|_0$ and then comparing the mean for each coordinate. Therefore, for both $\mathcal{A}_2^\rho$ and $\mathcal{A}_3^\alpha$, d_{ENTRY} is a tight characterization of the coordinate-level adversary. □

8.4. Proof of Theorem 2

Proof. Let $\text{disc}(\Sigma^{-1}) = \max_{x\in[-1,1]} \sqrt{x^T s(\Sigma^{-1})x}$ and let v be the vector with entries $(\Sigma^{-1}_{ii})^{-1/2}$. To complete the proof, we will show that $d^\infty_{\text{ENTRY}}(N(\mu, \Sigma), N(\mu + \rho v, \Sigma)) \leq \rho$. To do this, we are going to use a hybrid argument showing that by only hiding ρ fraction of the entries in the i-th coordinate, $N(\mu, \Sigma)$ and $N(\mu + \rho \mathbf{e}_i/\Sigma^{-1/2}_{ii}, \Sigma))$ become indistinguishable where $\mathbf{e}_i$ is the vector that has 1 in its i^{th} coordinate and 0 in the others. This is because, $d_{\text{TV}}(N(\mu, \Sigma), N(\mu + \rho\mathbf{e}_i/\Sigma^{-1/2}_{ii}, \Sigma))) \leq \rho$. By applying this argument sequentially for every coordinate, $N(\mu, \Sigma)$ and $N(\mu + \rho v, \Sigma)$ are indistinguishable under an $\mathcal{A}_2$ adversary. Since the total distance between μ and $\mu + \rho v$ in Mahalanobis

distance is at least $\rho \cdot \mathrm{disc}(\Sigma^{-1})$, the theorem follows. □

8.5. Proof of Corollary 1

Proof. We prove the following lemma that implies Corollary 1 when combined with Theorem 2.

Lemma 4. *For any $n \times n$ PSD matrix M, disc$(M) \in [\sqrt{n}, n]$*

We have that $s(M)$ is a PSD matrix with diagonal elements equal to 1. Consider a random x with uniformly random coordinates in $\{-1, 1\}$. Then, $\mathbb{E}[x^T s(M) x] = \mathrm{Trace}(s(M)) = n$. Thus, $\max_{x \in [-1,1]} \sqrt{x^T s(M) x} \geq \sqrt{n}$. This lower bound is tight for $M = I$.

For the upper-bound, we notice that since $s(M)$ is PSD, it holds that $|s(M)_{ij} + s(M)_{ji}| \leq 2$. To see this notice that $x^T s(M) x \geq 0$ for both $x = e_i + e_j$ and $x = e_i - e_j$.

Given this, we have that $x^T s(M) x \leq \frac{1}{2} \sum_{ij} |s(M)_{ij} + s(M)_{ji}| \leq n^2$. This gives the required upper-bound. Notice that the upper-bound is tight for the matrix M consisting entirely of 1's. □

8.6. Proof of Theorem 3

Proof. With a budget of α, $\mathcal{A}_3$ can concentrate its corruption on one particular coordinate, say the first coordinate. If $\alpha n \geq 1$, we will lose all information for the first coordinate, making mean estimation impossible. Since $\alpha < 1/n$, $\mathcal{A}_3$ can corrupt αn-fraction of first coordinates of all samples. Since the marginal distribution with respect to the first dimension is a univariate Gaussian, information-theoretically any mean estimator of the first coordinate must be $\Omega(\alpha n)$-far from the true mean of the first coordinate.

□

8.7. Proof of Theorem 4

Proof. First, we show the case for d^1_{ENTRY}.

$d^1_{\text{ENTRY}}(D_1, D_2) \leq d_{\text{TV}}(D_1, D_2)$ follows from

$$\begin{aligned} d^1_{\text{ENTRY}}(D_1, D_2) &= \inf_{\gamma \in \Gamma(D_1, D_2)} \frac{\| \mathbb{E}_{(x,y)\sim\gamma} [I(x,y)] \|_1}{n} \\ &= \inf_{\gamma \in \Gamma(D_1, D_2)} \mathop{\mathbb{E}}_{(x,y)\sim\gamma} \left[\frac{||x-y||_0}{n} \right] \\ &\leq \inf_{\gamma \in \Gamma(D_1, D_2)} \Pr_{(x,y)\sim\gamma} [x \neq y] \\ &= d_{\text{TV}}(D_1, D_2) \end{aligned}$$

Then we show that $d^1_{\text{ENTRY}}(D_1, D_2) \geq \frac{m_A}{n} d_{\text{TV}}(D_1, D_2)$.

We first show that for any $x \notin \ker(A)$, $||Ax||_0 \geq m_A$. Suppose by way of contradiction that $\Pi_i Ax$ is nonzero for fewer than m_A values of i. Call the rows of A $v_0^T, \ldots, v_{n-1}^T$ and let S be the subspace of $\mathbb{R}^r$ spanned by the v_i's. As $x \notin \ker(A)$, Ax is nonzero. Hence, $\langle x, v_i \rangle$ is nonzero for some i so $\Pi_S x$ is nonzero.

Now, let $\mathcal{B}$ be a basis for S containing $\Pi_S x$. Consider the subspace S' of S spanned by $\{v_i \mid \langle x, v_i \rangle = 0\}$. As $\Pi_{S'} x = 0$, $\Pi_S x$ cannot be an element of S' and so $\mathcal{B}$ is not a basis for S'. Thus, the dimension of S' is less than that of S; as $|\{v_i\}| - |\{v_i \mid \langle x, v_i \rangle = 0\}| < m_A$ we have a contradiction of the definition of m_A. Thus, if $x \neq 0 \in \mathbb{R}^r$, $\Pi_i Ax$ must be nonzero for at least m_A values of i, and hence $||Ax||_0 \geq m_A$.

Now, suppose that $(x, y) \sim \gamma$ for some $\gamma \in \Gamma(D_1, D_2)$. Then, $x = Ax'$ and $y = Ay'$ for some $x', y' \in \mathbb{R}^r$. If $x \neq y$, then $Ax' \neq Ay'$ so $x' - y' \notin \ker(A)$. Thus

$$||A(x' - y')||_0 \geq m_A$$

by the above, and so

$$\mathop{\mathbb{E}}_{(x,y)\sim\gamma} [||x-y||_0] \geq m_A \Pr_{(x,y)\sim\gamma} [x \neq y]$$

Therefore, we have that

$$\begin{aligned} d^1_{\text{ENTRY}}(D_1, D_2) &= \inf_{\gamma \in \Gamma(D_1, D_2)} \frac{\| \mathbb{E}_{(x,y)\sim\gamma} [I(x,y)] \|_1}{n} \\ &= \inf_{\gamma \in \Gamma(D_1, D_2)} \mathop{\mathbb{E}}_{(x,y)\sim\gamma} \left[\frac{||x-y||_0}{n} \right] \\ &\geq \inf_{\gamma \in \Gamma(D_1, D_2)} \frac{m_A}{n} \Pr_{(x,y)\sim\gamma} [x \neq y] \\ &= \frac{m_A}{n} d_{\text{TV}}(D_1, D_2) \end{aligned}$$

In the case of d^∞_{ENTRY}, the left hand side ($d^\infty_{\text{ENTRY}}(D_1, D_2) \geq \frac{m_A}{n} d_{\text{TV}}(D_1, D_2)$) follows from above by using the fact that $\|x\|_1 \leq n\|x\|_\infty$ for $x \in \mathbb{R}^n$. The right hand side follows from

$$\begin{aligned} d^\infty_{\text{ENTRY}}(D_1, D_2) &= \inf_{\gamma \in \Gamma(D_1, D_2)} \| \mathop{\mathbb{E}}_{(x,y)\sim\gamma} [I(x,y)] \|_\infty \\ &= \inf_{\gamma \in \Gamma(D_1, D_2)} \max_i \Pr_{(x,y)\sim\gamma} [x_i \neq y_i] \\ &\leq \inf_{\gamma \in \Gamma(D_1, D_2)} \Pr_{(x,y)\sim\gamma} [x \neq y] \\ &= d_{\text{TV}}(D_1, D_2) \end{aligned}$$

Therefore, the theorem holds for the d_{ENTRY} metric. □

8.8. Proof of Corollary 2

Proof. We can obtain the given upper bound relating the distance to d_{TV}. Since $d_{\text{TV}}(\mathcal{N}(0,1), \mathcal{N}(\mu,1)) = erf(\frac{\mu}{2\sqrt{2}})$, for small $\mu > 0$, $erf(\frac{\mu}{2\sqrt{2}}) = \Theta(\mu)$. Then

$$\begin{aligned} d_{\text{TV}}(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\mu',\Sigma)) &= d_{\text{TV}}(\mathcal{N}(0,I), \mathcal{N}(\Sigma^{-1/2}(\mu'-\mu), I)) \\ &= d_{\text{TV}}(\mathcal{N}(0,1), \mathcal{N}(\|\Sigma^{-1/2}(\mu'-\mu)\|_2, 1)) \\ &= d_{\text{TV}}(\mathcal{N}(0,1), \mathcal{N}(\|\mu'-\mu\|_\Sigma, 1)) = \Theta(\|\mu'-\mu\|_\Sigma) \end{aligned}$$

Applying Theorem 4, we get that $\|\mu - \mu'\|_\Sigma = O(\alpha \frac{n}{m_A})$. □

8.9. Proof of Theorem 5

Proof. We prove the theorem for both missing values and replaced values. In the case of missing values, for the lower bound, $\mathcal{A}^\alpha_3$ may corrupt at most $\frac{\alpha n}{m_A}$-fraction of the samples so that the coordinates are non-recoverable and shift part of the original distribution to anywhere along the axes of missing coordinates. Then the proof similarly follows the lower bound proof for estimating the mean of a Gaussian corrupted by $\mathcal{A}^\epsilon_1$. Hence, since we cannot distinguish between two Gaussians that share $1 - \frac{\alpha n}{m_A}$ of mass, $\|\hat{\mu} - \mu\|_\Sigma = \Omega(\alpha \frac{n}{m_A})$.

For $\mathcal{A}^\alpha_3$ that replaces values, we prove the following lemma and the theorem follows.

Lemma 5. *The adversary corrupts δ coordinates of a sample. Let $\tilde{x}$ be the corrupted sample and $x^* = Az^*$ be the original. We can only information-theoretically recover x^* from $\tilde{x}$ if and only if $\delta < \frac{m_A}{2}$. Furthermore, if $\delta < \frac{m_A}{2}$ then $\|\tilde{x} - Az^*\|_0 < \delta$ and $\|\tilde{x} - Az'\|_0 \geq m_A - \delta$ for any $z' \neq z^*$.*

Assume that if $\delta < \frac{m_A}{2}$ then $\|\tilde{x} - Az^*\|_0 < \delta$ and $\|\tilde{x} - Az'\|_0 \geq m_A - \delta$ for any $z' \neq z^*$. This implies that we can consider all possible subsets $I \subseteq \mathcal{U}$ where $|I| = n - \frac{m_A}{2}$ and solve the linear system of equations of $\tilde{x}_I = A_I z$ and output the solution z, which achieves smallest hamming distance to $\tilde{x}$, as z^*. If $\delta \geq \frac{m_A}{2}$, it is information theoretically impossible to recover x^* as $\arg\min_z \|\tilde{x} - Az\|_0$ may not be unique: since the corruptions are adversarial, z^* may not be part of the set of minimizers.

Assume $\delta < \frac{m_A}{2}$. Without loss of generality, let A be full rank. If not, the proof follows by replacing r with $\text{rank}(A)$ and considering the kernel of A. Let A_i denote the i-th row of matrix A and $\mathcal{U} = \{A_i : i \in [n]\}$. Let Δ denote the set of A_i's that correspond to the corrupted coordinates of $\tilde{x}$ so that $|\Delta| = \delta$. Define $\mathcal{S} \supseteq \Delta$ to be the smallest subset of $\mathcal{U}$ such that row space dimension (rank) of $A_{\mathcal{U}\setminus\mathcal{S}}$ is 1 less than that of A. By definition of m_A, $|\mathcal{S}| \geq m_A$.

The entries corresponding to rows $\mathcal{U} \setminus \mathcal{S}$ are uncorrupted, so if we solve the linear system $A_{\mathcal{U}\setminus\mathcal{S}} z = \tilde{x}_{\mathcal{U}\setminus\mathcal{S}}$, we will a get a 1-dimensional solution space for z. Thus, any z in this line will give at least $|\mathcal{U} \setminus \mathcal{S}|$ matching coordinates when multipied to A with x^*. Now, we can generate $|\mathcal{S}|$ many solutions, each corresponding to the solution to the linear system $A_{\mathcal{U}\setminus\mathcal{S}\cup\{s\}} z = \tilde{x}_{\mathcal{U}\setminus\mathcal{S}\cup\{s\}}$ for each $s \in \mathcal{S}$.

For s that corresponds to an uncorrupted entry in $\tilde{x}$, the solution to the linear system is the true solution z^* since none of the values in the system was corrupted. That gives us at least $|S| - \delta$ solutions out of $|S|$ solutions to be exactly z^*. Regardless of how the adversary corrupts the δ entries, if $\delta < \frac{m_A}{2}$, then the majority solution will always be z^* since $|S| - \delta > \frac{m_A}{2} > \delta$. Furthermore, for $z' \neq z^*$, z' can match at most $|\mathcal{U} \setminus \mathcal{S}| + \delta \leq n - m_A + \delta$ coordinates of $\tilde{x}$, i.e. $\|Az' - \tilde{x}\|_0 \geq m_A - \delta$. However, if $\delta \geq \frac{m_A}{2}$, then there is no clear majority so it is impossible to distinguish between the true solution and the other solution. In fact, when δ is strictly greater and corruptions adversarially chosen, $\|Az^* - \tilde{x}\|_0 = \delta$ and there exists some z', $\|Az' - \tilde{x}\|_0 = m_A - \delta < \|Az^* - \tilde{x}\|_0$.

□

8.10. Details of the Recovery Steps in the Algorithms in Section 4.1

The input is the structure matrix A and corrupted samples $\tilde{x}_i$ with M_i missing entries for $i = 1, 2, .., N$. When A is known, we iterate over all samples. If $M_i \geq m_A$, we discard $\tilde{x}_i$. Otherwise, we remove the missing entries in $\tilde{x}_i$ and get $\tilde{x}'_i$, and also remove the rows in A corresponding to those missing entries and get A'. Then we solve $A'z = \tilde{x}'_i$ and recover sample i with Az. When A is unknown, we perform matrix completion on $\tilde{X}$, the i^{th} row of which is $\tilde{x}'_i$. We first impute the missing entries with the coordinate-wise medians. Then we repeat the following procedure until convergence: 1) compute the rank-m_A projection of $\tilde{X}$ 2) replace the entries that are missing initially with the corresponding entries from the projection. The details of the projection and the completion procedure can be found in (Chunikhina et al., 2014).

8.11. Proof of Theorem 6

Proof. Define ϵ be the fraction of samples that has at least one corrupted coordinate. Note that the coordinate-level adversary must corrupt at least m_A coordinates of a sample to make his corruptions non-recoverable. Given that we can recover any sample with less than m_A corrupted coordinates, we have that $\epsilon \leq \frac{\alpha n}{m_A}$. If D is the original distribution on $\mathbb{R}^n$ and D' is the observed distribution, then $d_{\text{TV}}(D, D') \leq \epsilon$. Since d_{TV} between the two Gaussians is less than or equal to ϵ, the Tukey median algorithm achieves $\|\hat{\mu}_{\text{Tukey}} - \mu\|_\Sigma = O(\frac{\alpha n}{m_A})$. On the other hand, removing ϵ-fraction of a spherical Gaussian shifts the empirical mean by $O(\epsilon\sqrt{\log 1/\epsilon})$. Therefore, $\|\hat{\mu} - \mu\|_\Sigma = \tilde{O}(\frac{\alpha n}{m_A})$. □

8.12. Proof of Lemma 2

Proof. The adversary can simply hide one coordinate completely to prevent us from recovering that coordinate if at least one missing coordinate per sample in expectation is allowed. If the expected number of missing coordinates per sample is less than one, there must then be some positive fraction of samples with no missing coordinates; as we have infinite samples, we can select any $r + 1$ disjoint sets of $n - r$ such samples to satisfy the conditions in Lemma 1. □

8.13. Robust Mean Estimation Under Coordinate-fraction Adversaries with Unknown Structure

If we consider the case where the data is corrupted by $\mathcal{A}_3^\alpha$, we have the following result.

Theorem 9. *Assume samples $x_i = Az_i$ and z_i comes from a Gaussian such that $x_i \sim \mathcal{N}(\mu, \Sigma)$ with support in the range of A, but A is unknown. Under corruption $\mathcal{A}_3^\alpha$ with budget $\alpha < \frac{1}{n}$, recover missing coordinates by solving the matrix completion problem and discard any unrecoverable samples. The empirical mean $\hat{\mu}$ of the remaining samples satisfies $\|\mu - \hat{\mu}\|_\Sigma = \tilde{O}(\frac{\alpha n}{m_A})$, while the Tukey median $\hat{\mu}_{\text{Tukey}}$ of the remaining samples satisfies $\|\hat{\mu}_{\text{Tukey}} - \mu\|_\Sigma = O(\frac{\alpha n}{m_A})$.*

Theorem 9 is based on Theorem 6 and the Lemma 2.

8.14. Proof of Lemma 3

Proof. First, We introduce the concept of hidden patterns. The set of coordinates missing from a sample forms its hidden pattern. We only consider the hidden patterns which have been applied to infinitely many samples as if only finitely many samples share a pattern, the adversary could hide those samples completely with 0 budget.

When $\rho \geq \frac{m_A-1}{n}$, the adversary is able to hide $m_A - 1$ entries for every sample, and we cannot learn the structure from samples with only r visible entries.

When $\rho < \frac{m_A-1}{n}$, the adversary does not have enough budget to hide $m_A - 1$ entries of all the samples, so there exist some patterns with at least $r + 1$ coordinates visible. We use M to denote the number of such patterns, and p_l to denote the probability of the l^{th} pattern P_l, $l = 1, 2, \ldots, M$.

Next, we show a necessary condition for the adversary to prevent us from learning the structure. Since we have infinitely many samples, one group of $n - r$ samples satisfying the conditions in Lemma 1 is enough to learn the structure since we can find another r groups by choosing samples with the same hidden patterns.

It is obvious that the adversary has to hide at least one coordinate per pattern, otherwise we have infinitely many samples without corruption. No matter what the patterns the adversary provides, we try to get the samples satisfies the conditions in Lemma 1 by the following sampling procedure.

1. Start with one of the patterns, pick $r + 1$ visible coordinates of it to form the initial visible set V_1. Take one sample from this pattern to form the initial sample group G_0. Mark this pattern as checked.
2. For $K = 1, 2, 3, \ldots, M - 1$, take one of the unchecked patterns and check if it contains at least one visible coordinate not in V_K, the current visible set. If so, take one sample x_K from it and pick any one of its visible coordinates $v_K \notin V_K$. Add x_K to the sample group and v_K to the visible set: $G_{K+1} = G_K \cup \{x_K\}$, $V_{K+1} = V_K \cup \{v_K\}$. If not, skip it. Mark the pattern as checked.

We show by induction that any $k (k \leq K)$ different samples in G_K have at least $r + k$ coordinates in V_K not completely hidden. It is trivial that the property holds for $K = 1$. Assume that the property holds for K. According to the sampling procedure, when a new sample x_K comes, it has at least one visible coordinate v_K not in V_K. Consider any $k (k \leq K + 1)$ different samples in G_{K+1}. If the k samples don't include the new sample x_K, by the induction assumption they have at least $r + k$ coordinates in $V_K \subset V_{K+1}$ not completely hidden. If x_K is one of the k samples, again by the induction assumption the other $k - 1$ samples have at least $r + k - 1$ coordinates in V_K not completely hidden, plus v_K of x_K is also not hidden, so there are at least $r + k$ coordinates in V_{K+1} not completely hidden. Thus, the property also holds for $K + 1$. By induction, any k distinct samples from the group we get at the end of step 2 have at least $r + k$ coordinates not completely hidden, which means if the group has at least $n - r$ samples, the conditions in Lemma 1 can be satisfied.

Denote the set of the patterns being picked as $\mathcal{P}_P$ and the set of the patterns being skipped as $\mathcal{P}_S$. Based on the previous analysis, the adversary has to manipulate the patterns so that $|\mathcal{P}_P| \leq n - r - 1$, in which case the visible set cannot cover all the coordinates, which means there exists at least one common hidden coordinate for the patterns in $\mathcal{P}_S$ (otherwise the pattern where that coordinate is visible should have been picked). Since the fraction of hidden entries in that common coordinate is less than or equal to ρ, the sum of the probabilities of the patterns in $\mathcal{P}_S$ satisfies $\sum_{l:P_l \in \mathcal{P}_S} p_l \leq \rho$. Since all the patterns have at least one missing coordinate, we also have $p_{l:P_l \in \mathcal{P}_P} \leq \rho$. Thus, we have $\sum_{l=1}^{M} p_l \leq (n - r)\rho$. In such a case, the overall fraction of missing entries η satisfies

$$\begin{aligned}\eta &\geq \sum_{l=1}^{M} p_l \frac{1}{n} + (1 - \sum_{l=1}^{M} p_l)\frac{n-r}{n} \\ &\geq (n-r)\rho\frac{1}{n} + (1-(n-r)\rho)\frac{n-r}{n}\end{aligned}$$

The first inequality holds because for the samples with at least $r + 1$ visible entries, there are at least 1 missing entries per sample, and for the samples with less than $r + 1$ visible entries, there are at least $n - r$ missing entries per sample. In addition, η also satisfies $\eta \leq \rho$, so we have $\rho \geq \frac{n-r}{n+(n-r-1)(n-r)} = \frac{m_A-1}{n+(m_A-1)(m_A-2)}$, which is a necessary condition for the adversary. Thus, if $\rho < \frac{m_A-1}{n+(m_A-1)(m_A-2)}$, we can learn the structure and impute all the samples with at least r visible entries. □

8.15. Proof of Theorem 8

Proof. Algorithm 1 and the following analysis borrows significantly from the randomized approximation algorithm for the NP-hard problem Min-Unsatisfy in the work of (Berman & Karpinski, 2001).

If $\tilde{x}$ is in the subspace generated by A (i.e. there exists z such that $\tilde{x} = Az$), then the point $\tilde{x}$ must be either an uncorrupted point or a corrupt point where at least m_A coordinates corrupted. This is because m_A is equal to the minimum number of coordinates needed to move a point on the subspace generated by A to another point on the subspace.

If the point does not lie on the subspace, it is clearly a corrupted point. If this point is corrupted by more than $m_A/2$, it would be information-theoretically impossible to recover the true point as is shown in Lemma 5. However, changing this point to a different point on the subspace would not matter in the reduction to a robust mean estimation algorithm since it is a outlier either way.

Now assume this point is corrupted by at most $m_A/2$. By Lemma 2 and Theorem 1 of (Berman & Karpinski, 2001), the for loop results in the best $\tilde{z}$ such that $\|\tilde{x} - A\tilde{z}\|_0 \leq \frac{r}{c \ln r}\|\tilde{x} - x^*\|_0$ with high probability. In the case that the point is only corrupted by at most $\frac{cm_A \ln r}{2r}$ coordinates, then $\|\tilde{x} - A\tilde{z}\|_0 \leq \frac{m_A}{2}$. But the only point on the subspace such that it only needs at most $\frac{m_A}{2}$ coordinate changes to $\tilde{x}$ is x^*. Then this approximation algorithm performs exact recovery of x^* when at most $O(\frac{m_A \ln r}{r})$ coordinates are corrupted. Therefore, preprocessing points with this algorithm and then applying robust mean estimation yields a mean estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 = \tilde{O}(\frac{r}{\ln r} \cdot \frac{\alpha n}{m_A})$.

□

8.16. Connection between Recovery for Replacements and Sparse Recovery

We first show that the reduction from exact recovery for replacements to sparse recovery. Let $x^* = Az^*$ be the uncorrupted sample, and $\tilde{x}$ be the same sample with no more than δ coordinates corrupted. Let $e^* = \tilde{x} - x^*$, where $\|e^*\|_0 \leq \delta$, and we have $\tilde{x} = Az^* + e^*$. Take a non-trivial matrix $F \in \mathbb{R}^{p \times n}$ $(p < n)$ which satisfies $FA = 0$. Apply F to $\tilde{x}$ we have $y = F(A\tilde{x} + e^*) = Fe^*$. x^* can be recovered if we know what e^* is, so the recovery of x^* can be reduced to recovering e^* from y.

Candes & Tao (2005) show that when $\delta < \frac{m_A}{2}$, we can get the exact e^* by solving

$$\min_{e \in \mathbb{R}^n} \|e\|_0 \text{ subject to } Fe = y$$

Therefore, we reduce the problem of exact recovery for replacements to the problem of sparse recovery.

On the other hand, sparse recovery can also be reduced to exact recovery for replacements. Given the sparse recovery problem shown above, we take $A \in \mathbb{R}^{n \times r}$ such that the columns of A span the null space of F, and $\tilde{x} \in \mathbb{R}^n$ such that $F\tilde{x} = y$. For any e satisfying $Fe = y$, we have $F\tilde{x} = Fe$, and therefore $\tilde{x} = e + Az$ for some $z \in \mathbb{R}^r$. In addition, for any $z \in \mathbb{R}^r$, we can find $e = \tilde{x} - Az$ satisfying $Fe = y$. Thus, solving the exact recovery problem $\min_{z \in \mathbb{R}^r} \|\tilde{x} - Az\|_0$ will also give the sparsest e.

8.17. Computationally Efficient Algorithms for Sparse Recovery

Basis pursuit (BP) (Candes & Tao, 2005) and orthogonal matching pursuit (OMP) (Davenport & Wakin, 2010) are computationally efficient algorithms that get the exact e^* defined in Section 8.16 under certain conditions.

BP approximates the sparse recovery problem by

$$\min_{e \in \mathbb{R}^n} \|e\|_1 \text{ subject to } Fe = y \tag{1}$$

which is convex and can be solved by linear programming.

OMP is a greedy algorithm that gives a solution by the following procedure:

- **Step 1** Initialize the residual $r^0 = y$, the index set $\Lambda^0 = \emptyset$, and the iteration counter $l = 0$.
- **Step 2** Find the column of F that has the largest inner product with the current residual and add its index to the index set: $\Lambda^{l+1} = \Lambda^l \cup \{\arg\max_i |F_i^T r^l|\}$.

- **Step 3** Update the estimation and the residual: $e^{l+1} = \arg\min_{v:\text{supp}\,(v)\subseteq\Lambda^{l+1}} \|y - Fv\|_2$, $r^{l+1} = y - Fe^{l+1}$.
- **Step 4** Output e^{l+1} if converged, otherwise increment l and return to Step 2.

RIP-based Guarantee for BP and OMP Candes & Tao (2005) introduce the Restricted Isometry Property (RIP) that characterizes the orthonormality of matrices when operating on sparse vectors.

Definition 4. *(RIP) A matrix F satisfies the RIP of order k if*

$$(1-\zeta)\|c\|^2 \leq \|Fc\|^2 \leq (1+\zeta)\|c\|^2 \tag{2}$$

for all real coefficients c with $\|c\|_0 \leq k$. $\zeta \in (0,1)$ is a constant.

Based on the RIP condition, Cai & Zhang (2013) and Davenport & Wakin (2010) show that BP and OMP recover a K-sparse e exactly when F satisfies certain RIP conditions, and we state their results here as the following lemma.

Lemma 6. *Consider the problem of recovering a K-sparse e^* from Fe^*. BP recovers e^* exactly if F satisfies the RIP of order K with $\zeta < \frac{1}{3}$; OMP recovers e^* exactly in K iterations if F satisfies the RIP of order $K+1$ with $\zeta < \frac{1}{3\sqrt{K}}$.*

Matrices Satisfying RIP Gaussian matrices $F \in \mathbb{R}^{p\times n}$ whose entries are independent realizations of $\mathcal{N}(0, \frac{1}{p})$ satisfy RIP with certain probability, which can be shown by the following lemma that comes from Baraniuk et al. (2008).

Lemma 7. *Consider a Gaussian matrix $F \in \mathbb{R}^{p\times n}$ whose entries are independent realizations of $\mathcal{N}(0, \frac{1}{p})$. For a given $0 < \zeta < 1$, there exist constants $c_1, c_2 > 0$ such that the RIP of order k ($k \leq c_1 p/log(n/k)$) with ζ holds for Gaussian metrics F with probability at least $1-2e^{-c_2 p}$. c_1, c_2 have the relation that $c_2 = c_0(\zeta/2) - c_1[1+(1+\log(12/\zeta))/\log(n/k)]$, where $c_0(x) = x^2/4 - x^3/6$ is a function.*

Note that one can choose sufficiently small c_1 so that $c_2 > 0$.

Recovery for Gaussian A When the structure A is a Gaussian matrix, we can perform efficient recovery by BP or OMP, with the following guarantee:

Theorem 10. *Suppose $x^* = Az^*$ is an uncorrupted sample, where $A \in \mathbb{R}^{n\times r}$ is a Gaussian matrix whose entries are i.i.d. realizations of some Gaussian random variable $\mathcal{N}(0, \sigma^2)$. $\tilde{x} = x^* + e^*$ is the corrupted version of x^*, and e^* is K-sparse with $K < m_A/2$. x^* can be recovered exactly from $\tilde{x}$ with probability at least $1 - 2e^{-c_2 p}$ by the following procedure:*

- ***Step 1:*** *Let $k = K$. Choose $\zeta < \frac{1}{3}$, and c_1 so that $c_2 = c_0(\zeta/2) - c_1[1 + (1 + \log(12/\zeta))/\log(n/k)] > 0$. Choose p such that $k \leq c_1 p/log(n/k)$.*
- ***Step 2:*** *For i from 1 to p, randomly sample a vector v_i from the unit sphere in the null space of A^T, and a scalar u_i from the chi-square distribution χ^2_n of degree n. Construct $F \in \mathbb{R}^{p\times n}$ where the i^{th} row is $\sqrt{\frac{1}{p}u_i}v_i^T$.*
- ***Step 3:*** *Solve $Fe = F\tilde{x}$ by BP and denote the result as $\hat{e}$. Output $\hat{x} = \tilde{x} - \hat{e}$.*

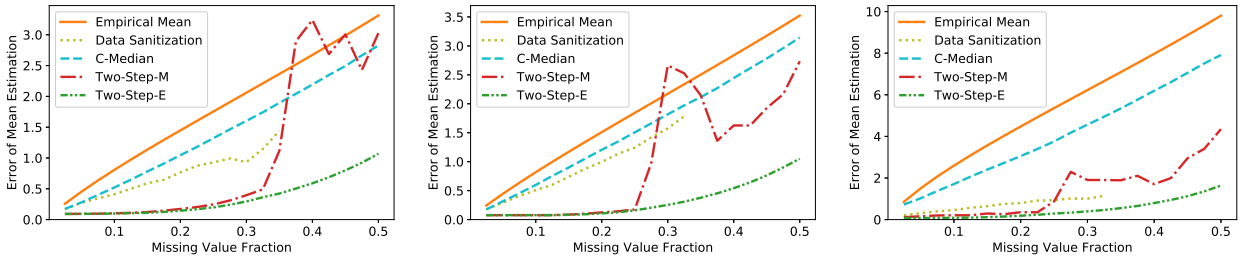
Proof. Since each row of F is sampled from the null space of A^T, we have $FA = 0$ and $F\tilde{x} = Fe$. Because A is a Gaussian matrix, v_i is a random unit vector sampled from the n dimension sphere, and each entry of it can be considered as i.i.d. samples from $\mathcal{N}(0, \frac{1}{p})$ after we multiply it by $\sqrt{\frac{1}{p}u_i}$. Hence, F is a Gaussian matrix whose entries are independent realizations of $\mathcal{N}(0, \frac{1}{p})$. The satisfaction of the RIP condition and the exact recovery property follow from Lemma 7 and 6 respectively. Therefore, with probability at least $1 - 2e^{-c_2 p}$, $\hat{e} = e^*$ and $\hat{x} = x^*$. □

Similar result for OMP can be derived by simply replacing the conditions for k and ζ in Step 1 by $k = K+1$ and $\zeta < \frac{1}{3\sqrt{K}}$.

8.18. Detailed Experimental Evaluation

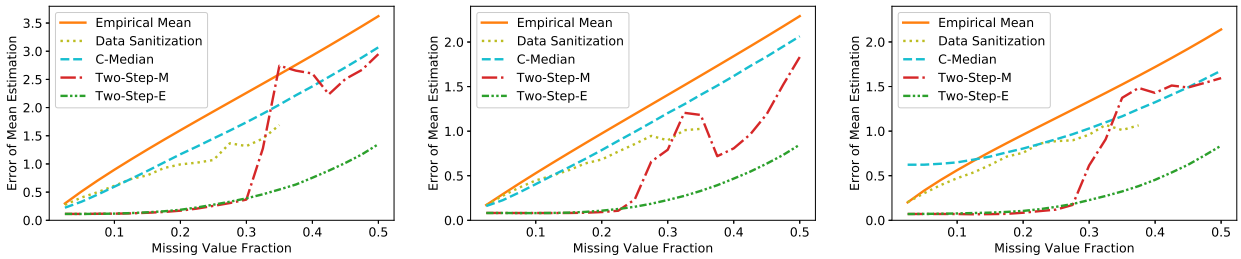
We show a detailed description of our experiments. We consider both real-world data that may not exhibit linear structure and synthetic data that does not always follow a Gaussian distribution.

Methods and Experimental Setup We consider the following mean estimation methods:



(a) Gaussian z (b) Uniform z (c) Exponential z

Figure 3. Mean estimation error (Mahalanobis) for synthetic data sets.



(a) Gaussian z (b) Uniform z (c) Exponential z

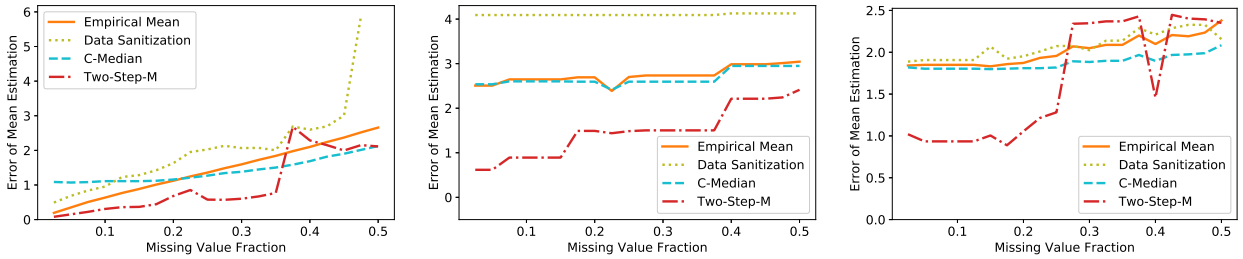
Figure 4. Mean estimation error (in l_2) for synthetic data sets.

- **Empirical Mean:** Take the mean for each coordinate, ignoring all missing entries.
- **Data Sanitization:** Remove any samples with missing entries, and then take the mean of the rest of the data.
- **Coordinate-wise Median (C-Median):** Take the median for each coordinate, ignoring all missing entries.
- **Our Method with Matrix Completion (Two-Step-M):** Use iterative hard-thresholded SVD (ITHSVD) (Chunikhina et al., 2014) to impute the missing entries. Take the mean afterwards. We use randomized SVD (Halko et al., 2011) to accelerate.
- **Out Method with Exact Recovery (Two-Step-E):** For each sample, build a linear system based on the structure and solve it. If the linear system is under-determined, do nothing. Then, take the mean while ignoring the remaining missing values.

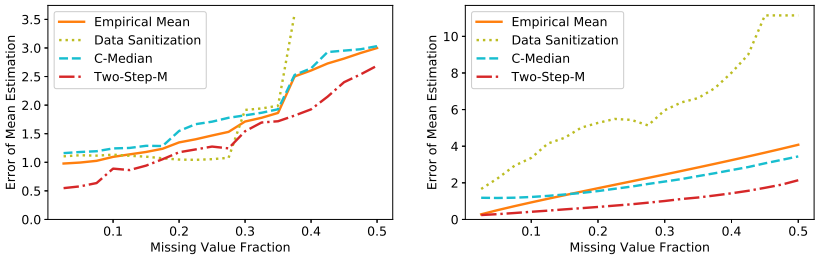
The methods can be classified into three categories, based on the amount of structural information they leverage: (1) Empirical Mean, Data Sanitization, and C-Median ignore the structure information; (2) Two-Step-M assumes there exists some unknown structure but it can be inferred from the visible data; (3) Two-Step-E knows exactly what the structure is and uses it to impute the missing values.

In each experiment presented below, we inject missing values by hiding the smallest ϵ fraction of each dimension. For synthetic data sets, the true mean is derived from the data generation procedure. For real-world data sets, we use the empirical mean of the samples before corruption approximate the true mean. For synthetic data sets, we consider the l_2 and Mahalanobis distances to measure the estimation accuracy of different methods. For real-world data, we only consider the l_2 distance between the estimated mean and the true empirical mean of the data before corruption.

Mean Estimation on Synthetic Data We show that redundancy in the corrupted data can help improve the robustness of mean estimation. We test all the methods on synthetic data sets with linear structure ($x = Az$) and three kinds of latent variables (z): 1) Gaussian, 2) Uniform, and 3) Exponential. Each sample x_i is generated by $x_i = Az_i$, where z_i is sampled from the distribution D_z describing the latent variable z. We set A to be a diagonal block matrix with two 8×4 blocks



(a) Leaf (b) Breast Cancer Wisconsin (c) Blood Transfusion



(d) Wearable Sensor (e) Mice Protein

Figure 5. Error of mean estimation on real-world data sets.

Table 2. Properties of the real-world data sets in our experiments.

Data Set	Samples	Features	ITHSVD Rank
Leaf	340	14	3
Breast Cancer	69	10	3
Blood Transfusion	748	5	3
Wearable Sensor	52081	9	4
Mice Protein Expr.	1080	77	10

generated randomly and fixed through the experiments. In every experiment, we consider a sample with 1,000 data vectors. To reduce the effect of random fluctuations, we repeat our experiments for five random instances of the latent distribution D_z for each type of latent distribution and take the average error.

The results for the above experiments are shown in Figure 3. This figure shows the mean estimation error of different methods measured using the Mahalanobis distance. Additional results with the l_2 distance are shown in Figure 4. We see that estimators that leverage the redundancy in the observed data to counteract corruption yield more accurate mean estimates. This behavior is consistent across all types of distributions and not only for the case of Gaussian distributions that the theoretical analysis in Section 3 focuses on. We see that the performance of Two-Step-M (when the structure of A is considered unknown) is the same as that of Two-Step-E (when the structure of A is known) when the fraction of missing entries is below a certain threshold. Following our analysis in Section 4, this threshold corresponds to the conditions for which the subspace spanned by the samples can be learned from the visible data. Finally, we point out that we do not report results for Data Sanitization when the missing fraction is high because all samples get filtered.

Mean Estimation on Real-world Data We turn our attention to settings with real-world data with unknown structure. We use five data sets from the UCI repository (Dua & Graff, 2017) for the experiments in this section. Specifically, we consider: Leaf (Silva et al., 2013), Breast Cancer Wisconsin (Mangasarian & Wolberg, 1990), Blood Transfusion (Yeh et al., 2009), Wearable Sensor (Torres et al., 2013), and Mice Protein Expression (Higuera et al., 2015). For each data set, we consider the numeric features; all of these features are also standardized. For all the data sets, we report the l_2 error. We summarize the size of the data sets along with the rank used for Two-Step-M in Table 2. As the structure is unknown, we omit Two-Step-E. We show the results in Figure 5. We find that Two-Step-M always outperforms Empirical Mean and C-Median on three data

sets (Breast Cancer Wisconsin, Wearable Sensor, Mice Protein Expression). For the other two (Leaf, Blood Transfusion), the error of Two-Step-M can be as much as two-times lower than the error of the other two methods for small ϵ's ($\epsilon < 0.35$ for Leaf and $\epsilon < 0.25$ for Blood Transfusion). We also see that the estimation error becomes very high only for large values of ϵ. It is also interesting to observe that Data Sanitization performs worse than the Empirical Mean and the C-Median for real-world data. Recall that the opposite behavior was recorded for the synthetic data setups in the previous section. Overall, these results demonstrate that structure-aware robust estimators can outperform the standard filtering-based robust mean estimators even in setups that do not follow the linear structure setup in Section 3.