A Combinatorial Approach to Robust PCA

Weihao Kong ✉
Google Research, Mountain View, CA, USA

Mingda Qiao ✉ (ORCID)
University of California, Berkeley, CA, USA

Rajat Sen ✉ (ORCID)
Google Research, Mountain View, CA,USA

──── **Abstract** ────

We study the problem of recovering Gaussian data under adversarial corruptions when the noises are low-rank and the corruptions are on the coordinate level. Concretely, we assume that the Gaussian noises lie in an unknown k-dimensional subspace $U \subseteq \mathbb{R}^d$, and s randomly chosen coordinates of each data point fall into the control of an adversary. This setting models the scenario of learning from high-dimensional yet structured data that are transmitted through a highly-noisy channel, so that the data points are unlikely to be entirely clean.

Our main result is an efficient algorithm that, when $ks^2 = O(d)$, recovers *every single* data point up to a nearly-optimal ℓ_1 error of $\tilde{O}(ks/d)$ in expectation. At the core of our proof is a new analysis of the well-known Basis Pursuit (BP) method for recovering a sparse signal, which is known to succeed under additional assumptions (e.g., incoherence or the restricted isometry property) on the underlying subspace U. In contrast, we present a novel approach via studying a natural combinatorial problem and show that, over the randomness in the support of the sparse signal, a high-probability error bound is possible even if the subspace U is *arbitrary*.

1 Introduction

The flurry of recent work on algorithmic robust statistics studies various settings of learning and estimation in the presence of adversarial outliers (see e.g., [12] for a survey). Techniques have been developed for designing algorithms that are provably correct and both sample- and computationally-efficient. In most of these settings, however, an important assumption is that the majority of the data points are clean, so that non-trivial estimation is possible at least information-theoretically. Alternative models such as list-decodable learning ([4, 7]), while allowing a majority of the data to be corrupted, still require the clean data to constitute a non-negligible portion of the dataset.

This work explores a setting of robust learning from high-dimensional and structured data that are densely corrupted. Concretely, for some $d \gg k$, each clean data point x is a d-dimensional vector lying in a k-dimensional subspace $U \subseteq \mathbb{R}^d$. When d is huge, each data point is no longer accurately modeled by an atomic object that is either entirely clean or fully under the control of the adversary. For instance, when the data are transmitted through a noisy channel, it is conceivable that only a small fraction of the data *entries* get corrupted. Even if each entry gets corrupted independently with a small probability ϵ, the fraction of entirely clean data points, $(1 - \epsilon)^d$, could still be exponentially small. Indeed, we will focus on the extreme case that *every* data point may contain s corrupted entries, for some $s \ll d$.

Unfortunately, even in a seemingly benign setup where both k and s are small, an all-powerful adversary might still prevent us from extracting any information from the data. For instance, when the subspace U consists of the vectors supported on the first k coordinates, an adversary that corrupts $s \geq k$ entries per sample can easily wipe out all the information. However, not all hope is lost – in this example, the critical reason for the adversary to succeed is that the corruptions are perfectly aligned with the informative entries in the data. In many practical scenarios (e.g., when the corruptions result from the noisy transmission channel), it is reasonable to assume that the *locations* of corruptions are independent of the structure of the subspace U, so the worst-case scenario above is unlikely to happen. Concretely, we will assume in this work that the corruption on each data point is restricted to a randomly chosen subset of s coordinates.

The model that we informally defined above is closely related to the vast body of literature on *sparse recovery* and *robust PCA*. When the clean data point lies in a *known* subspace in our setup, the problem is equivalent to recovering the (additive) corruption, which is s-sparse, given a linear measurement of it; see Remark 1 for details. This is exactly the problem of sparse recovery, though we note that most previous work focused on the structures of the linear measurements that enable accurate and efficient recovery, whereas we consider an alternative setup where the linear measurement is essentially arbitrary, but the support of the sparse vector is uniformly distributed. If we arrange the observed data points into a matrix, the problem can be rephrased as decomposing this matrix into the sum of a low-rank matrix (consisting of the clean data) and a sparse matrix (consisting of the additive corruptions). This is known as robust PCA, and prior work has identified various suites of assumptions – on both the incoherence of the low-rank component and the support of the sparse component – under which an accurate decomposition can be efficiently computed. We discuss the related work on these two problems in more detail in Section 3.

1.1   Problem Setup

Notations

$[n]$ is the set $\{1, 2, \ldots, n\}$. We denote by $\binom{S}{s}$ the family of all size-s subsets of a set S. For vector v, $\|v\|_p$ is its p-norm when $p \geq 1$, and $\|v\|_0$ is the number of non-zero entries in v. For $v \in \mathbb{R}^d$ and $S \subseteq [d]$, $v|_S \in \mathbb{R}^{|S|}$ is the restriction of v to the coordinates with indices in S. Similarly, for a matrix A with d columns, $A|_S$ is the restriction of A to the columns indexed by $S \subseteq [d]$. The $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ notations hide polylog(d) factors, where d is the ambient dimension in the problem.

We first formally state the problem of recovering a vector lying in a *known* low-dimensional subspace when several randomly chosen entries are corrupted. This will be the key step in the recovery of an entire dataset that lie in an unknown subspace, a problem that we define later.

▶ **Problem 1** (Recovery of One Data Point)**.** *We are given a k-dimensional subspace $U \subseteq \mathbb{R}^d$ in which the unknown x lies. Set S is chosen uniformly at random from $\binom{[d]}{s}$. The adversary produces the corrupted data point $\tilde{x}$ subject to:*

▬ *Corruptions are restricted to S: $\tilde{x}_i = x_i$ for every $i \in [d] \setminus S$.*
▬ *Corruptions are bounded: $\|\tilde{x} - x\|_\infty \leq B$.*
Our goal is to output an estimate $\hat{x}$ of the clean data x with a small error.

Note that we may assume $B = 1$; otherwise we simply scale everything down by a factor of B. Also note that another natural corruption model – in which each coordinate is independently included in S with probability ϵ – corresponds to the case that $s \sim \mathsf{Binomial}(d, \epsilon)$

in Problem 1. Later, we will analyze an algorithm for Problem 1 that is oblivious to the value of s, so the guarantee for such an algorithm immediately implies a guarantee under the independent corruption setup.

The boundedness assumption $\|\tilde{x} - x\|_\infty \leq B$ in Problem 1 arises naturally in the context of the following problem of recovering a Gaussian dataset. In this setup, we can easily obtain a crude estimate for each coordinate of the mean vector, and then safely disregard the data entries that are far from the estimated mean.

▶ **Problem 2** (Recovery of Structured Data under Coordinate-Level Corruption)**.** *The clean data $x^{(1)}$, $x^{(2)}$, ..., $x^{(n)}$ are independent samples from an unknown Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, where $\mathrm{rank}(\Sigma) = k$. Sets $S^{(1)}, \ldots, S^{(n)}$ are chosen independently and uniformly at random from $\binom{[d]}{s}$. The adversary produces the corrupted data $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(n)}$ such that $\tilde{x}_j^{(i)} = x_j^{(i)}$ for every $j \in [d] \setminus S^{(i)}$. Our goal is to recover the clean data given the corrupted data.*

Note that a solution to Problem 2 immediately allows us to post-process the recovered data and extract information about the underlying distribution (e.g., the mean μ).

1.2 Our Results

We analyze the natural ℓ_1 convex relaxation, termed the Basis Pursuit (BP) method by Chen, Donoho and Saunders [8], for Problem 1: Given the corrupted data point $\tilde{x}$ and subspace U, we solve the following optimization problem to obtain an estimate of x:

$$\text{minimize} \quad \|\hat{x} - \tilde{x}\|_1$$
$$\text{subject to} \quad \hat{x} \in U \tag{1}$$

Note that the boundedness parameter B and the sparsity s of the corruption are not used in (1).

▶ **Remark 1.** Basis pursuit is more often formulated as minimizing $\|x\|_1$ subject to $Ax = y$, where x is a sparse signal, A specifies a linear measurement, and y is the observation from the measurement. The optimization above is equivalent to this more usual formulation: we may write $\delta = \tilde{x} - \hat{x}$ and let A be the $(d-k) \times d$ matrix with row space $U^\perp$, the orthogonal complement of U. The constraint $\hat{x} \in U$ in (1) is equivalent to $A(\tilde{x} - \delta) = 0$. Thus, Program (1) is equivalent to the minimization of $\|\delta\|_1$ subject to $A\delta = A\tilde{x}$.

Our main result states that the BP program in (1), while being a simple and natural relaxation of the sparsity objective, approximately denoises the data with high probability, even if no assumptions are made on the subspace U.

▶ **Theorem 2.** *In the setup of Problem 1 with $B = 1$, let x^* be an optimum of the BP method. For any $t > 0$, over the randomness in the set of corrupted coordinates, $\Pr\left[\|x^* - x\|_1 \geq t\right]$ is upper bounded by both*

$$\frac{1}{(\lfloor t/4 \rfloor + 1)!} \cdot \left(\frac{12 s^2 k}{d}\right)^{\lfloor t/4 \rfloor + 1}$$

and

$$\frac{24 k s}{d}.$$

Furthermore, assuming $s \leq \sqrt{d/2}$, the probability is also upper bounded by

$$12 k \cdot \left(\frac{2s}{d}\right)^{1 + \lfloor t/(48k) \rfloor}.$$

The second bound above holds uniformly for all $t > 0$. The third bound is sharper for larger values of t: assuming $s = O(\sqrt{d})$, the probability of incurring error $\geq t$ is at most $O(k) \cdot (2s/d)^{\Omega(t/k)}$, which decays exponentially as t grows. The first bound, in comparison, removes the $1/k$ factor in the exponent at the cost of a slightly larger base in the exponentiation.

An easy calculation (deferred to the full version) gives an upper bound on the expected ℓ_1 error.

▶ **Corollary 3** (Expected error). *In the setup of Problem 1, the expected ℓ_1 error of the BP method is $\tilde{O}\left(\frac{Bks}{d} \cdot \max\left\{\frac{ks^2}{d}, 1\right\}\right)$.*

Assuming $ks^2 = O(d)$, the error bound in Corollary 3 reduces to $\tilde{O}(Bks/d)$, which cannot be improved in general – the expected ℓ_1 error is at least $\Omega(Bks/d)$ when the subspace is $U = \text{Span}(\{e_1, e_2, \ldots, e_k\})$. This is because, in expectation, ks/d out of the first k coordinates get corrupted, and there is no hope in recovering each corrupted entry with an error better than $\Omega(B)$.

In comparison, the naïve algorithm that simply outputs $\tilde{x}$ gives an error of $\Omega(Bs)$ in the worst case, which is much worse in the low dimensional ($k \ll d$) regime. Another straightforward algorithm is based on the observation that, since U is k-dimensional, there exist k coordinates that uniquely determine a vector in U. With probability $\geq 1 - \frac{ks}{d}$, none of them gets corrupted and x can be perfectly recovered by solving a linear system. However, when at least one of the k coordinates does get corrupted, it is unclear that the recovery satisfies a reasonable error bound.

The setup of Problem 2 differs from the sparse recovery problem in two aspects: (1) We don't know the subspace that contains the clean data points; (2) The corruptions are no longer bounded. Our next theorem shows that, in the $ks = O(d)$ regime, the subspace can be efficiently recovered.

▶ **Theorem 4.** *In the setup of Problem 2, let U be the column space of Σ. Assuming $ks \leq c_0 \cdot d$ for sufficiently small constant $c_0 > 0$, there is an algorithm that recovers the subspace $U' := \text{Span}(U \cup \{\mu\})$ with high probability using $n = \tilde{O}(k^2)$ samples and in $\text{poly}(d)$ time.*

The theorem above implicitly assumes that we can access the samples as well as perform arithmetic operations with unlimited accuracy. To translate this result to more realistic models of computation (with limited precision), we note that the bottleneck in our algorithm will be solving an instance of robust linear regression in which the marginal distribution is a Gaussian with the covariance matrix being a full-rank principal submatrix of Σ. Thus, the algorithm will likely be numerically stable as long as every invertible principal submatrix of Σ is reasonably well-conditioned (compared to the bit precision). See the full version for a more detailed discussion.

Theorems 2 and 4 together give the following guarantee for recovering an entire Gaussian dataset.

▶ **Theorem 5.** *In the setup of Problem 2, assuming $n = \tilde{\Omega}(k^2)$ and $ks \leq c_0 \cdot d$ for some sufficiently small constant $c_0 > 0$, there is an algorithm that outputs n estimates $\hat{x}^{(1)}, \ldots, \hat{x}^{(n)}$ in $\text{poly}(d, n)$ time, such that each $\|\hat{x}^{(i)} - x^{(i)}\|_1$ is bounded by $\tilde{O}\left(\frac{Bks}{d} \cdot \max\left\{\frac{ks^2}{d}, 1\right\}\right)$ in expectation, where $B := \sqrt{\max_{i \in [d]} \Sigma_{ii}}$.*

As a direct corollary, the recovered data points allow us to estimate the unknown mean μ of the underlying Gaussian distribution.

▶ **Corollary 6.** *Under the same assumptions as in Theorem 5, there is an efficient algorithm that outputs $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_1$ is $\tilde{O}\left(\frac{Bks}{d} \cdot \max\left\{\frac{ks^2}{d}, 1\right\} + \frac{Bd}{\sqrt{n}}\right)$ in expectation.*

The additional error term $Bd/\sqrt{n}$ is unavoidable, even when $k = 1$ and no corruption is present. When $ks^2 = O(d)$ and $n = \Omega(d^4/k^2s^2)$, the error bound in Corollary 6 reduces to $\tilde{O}(Bks/d)$. Note that the naïve approach of estimating each coordinate of the mean vector separately would give a worse error guarantee when $k \ll d$. Indeed, when restricted to the i-th coordinate, the dataset becomes samples drawn from $\mathcal{N}(\mu_i, \Sigma_{ii})$, in which an $O(s/d)$ fraction gets corrupted. We can at best estimate μ_i up to an additive error of $O(\sqrt{\Sigma_{ii}} \cdot s/d) = O(Bs/d)$, so the ℓ_1 error bound would be $O(Bs)$, i.e., larger by a factor of d/k. In light of this comparison, Corollary 6 formalizes the intuition that since the problem is essentially k-dimensional, we only need to pay the $O(Bs/d)$ cost (the error in the one-dimensional case) k times instead of d times.

2   An Overview of Our Approach

We sketch the proofs of Theorems 2, 4 and 5 here. Our analysis of the Basis Pursuit method for Problem 1 is inspired by the Restricted Nullspace Property (RNP), which is known to imply the exact recovery guarantee of BP. We introduce a quantitative version of RNP (Lemma 9) that gives a necessary condition – on both the set S of corrupted coordinates and the subspace U – for BP to incur a large error in its recovery.

The technical highlight of our analysis is Lemma 10, which gives a combinatorial constraint on the family of size-s subsets of $[d]$ on which the aforementioned condition holds. Crucially, this constraint is a consequence of only the dimensionality of U, and none of the usual structural assumptions (such as incoherence or restricted isometry) on U. We then sketch how this constraint implies the desired high-probability error bound via studying a combinatorial problem similar to those in the extremal combinatorics literature.

Then, we explain how Problem 2 is solved as an application of the above. To recover the unknown subspace in the $ks = O(d)$ regime (Theorem 4), the key observation is that the recovery problem is essentially $(k + 1)$-dimensional, so the known results on robust linear regression (against sample-level corruptions) can be applied. Theorem 5 then follows from the two earlier theorems and a simple reduction of the problem to the bounded-corruption case.

2.1   A New Analysis of Basis Pursuit

Generalization of the RNP

With respect to subspace $U \subseteq \mathbb{R}^d$, we say that $S \subseteq [d]$ is *t-dominant* for $t > 0$, if there exists $u \in U$ such that $\|u\|_1 = 1$ and

$$\sum_{i \in S} |u_i| \cdot \mathbb{1}\left[|u_i| \leq 1/t\right] \geq 1/2.$$

We will show in Lemma 9 that the BP method incurs an ℓ_1 error of t only if the set S of corrupted coordinates is t-dominant.

The definition above becomes more natural in the context of the Restricted Nullspace Property, on which the textbook recovery guarantee of BP (e.g., [18]) relies.[1]

---

[1] As mentioned in Remark 1, subspace U in Problem 1 corresponds to the nullspace of the design matrix A in the usual formulation of BP, so Definition 7 is indeed about a nullspace.

▶ **Definition 7** (Restricted Nullspace Property). *A linear subspace $U \subseteq \mathbb{R}^d$ satisfies s-RNP if it holds for every $u \in U \setminus \{0\}$ and $S \in \binom{[d]}{s}$ that $\|u|_S\|_1 < \|u|_{[d] \setminus S}\|_1$.*

s-RNP is easily seen to be equivalent to the non-existence of 1-dominant sets of size s – when $t = 1$, the condition $\mathbb{1}\left[|u_i| \leq 1/t\right]$ in the definition of t-dominant sets is implied by $\|u\|_1 = 1$.

It is well-known that RNP is equivalent to the exact recovery guarantee of the Basis Pursuit method.

▶ **Theorem 8** (Theorem 7.8 of [41]). *In the setup of Problem 1, U satisfies s-RNP if and only if BP recovers every $x \in U$ with zero error for any $S \in \binom{[d]}{s}$.*

In particular, the BP method may err when the RNP is violated. For example, the subspace $U = \text{Span}(\{e_1, e_2, \ldots, e_k\})$ does not satisfy s-RNP for any $s \geq 1$, since $[s]$ is a 1-dominant set witnessed by $u = e_1$. As discussed earlier, exact recovery in this setup is not guaranteed, since whenever $S \cap [k]$ is non-empty, the adversary may arbitrarily corrupt one of the k informative coordinates.

On the bright side, when S is uniformly drawn from $\binom{[d]}{s}$, BP still satisfies a high-probability error bound on this example. We can easily verify that for this choice of U, the BP method simply zeros out the last $d - k$ coordinates of $\tilde{x}$, and the ℓ_1 error is linear in $|S \cap [k]|$, which is ks/d in expectation and has a sub-exponential tail.

This high-probability error bound can be alternatively explained in terms of of t-dominant sets. It can be easily verified that for $t \leq k$, each fixed $S \in \binom{[d]}{s}$ is t-dominant w.r.t. U if and only if $|S \cap [k]| \geq t/2$. When S is chosen uniformly at random, because of the sub-exponential tail of $|S \cap [k]|$, the probability of $|S \cap [k]| \geq t/2$ is exponentially small in t. Lemma 9 would then imply a high-probability error bound of BP. Perhaps surprisingly, our Theorem 2 gives quantitatively similar bounds for every k-dimensional subspace U. In other words, the simple example above is essentially the worst case.

A constraint on dominant sets

The crux of our analysis is then to upper bound the number of t-dominant sets w.r.t. the subspace U. As a warmup, we start with the special case that $t = 1$, and assume that the dominance is always *strict*, in the sense that every 1-dominant set $S \in \binom{[d]}{s}$ has a witness $u \in U$ such that $\|u\|_1 = 1$ and $\|u|_S\|_1 > 1/2$ (instead of $\|u|_S\|_1 \geq 1/2$).

Let $\mathcal{F}$ be the family of strictly 1-dominant sets. We note the following property of $\mathcal{F}$:

The family $\mathcal{F}$ does not contain $k + 1$ sets that are pairwise disjoint.

To see this, suppose towards a contradiction that $S_1, S_2, \ldots, S_{k+1} \in \mathcal{F}$ are pairwise disjoint, and vector $u^{(i)} \in U$ witnesses the strict dominance of S_i. Consider the $(k+1) \times d$ matrix $A = \begin{bmatrix} u^{(1)} & u^{(2)} & \cdots & u^{(k+1)} \end{bmatrix}^\top$. We obtain another $(k+1) \times (k+1)$ matrix A' by appropriately aggregating the columns of A. First, if $j \in S_i$ and $A_{ij} < 0$, we negate the j-th column of A. Then, we define the i-th column of A' as the sum of the j-th column of A over $j \in S_i$. For instance, for $(S_1, S_2, S_3) = (\{1, 2\}, \{3, 4\}, \{5\})$ and

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & 2 \\ 3 & 3 & -4 & 5 & 2 \\ 1 & -3 & 0 & 2 & -7 \end{bmatrix},$$

Columns 2, 3, and 5 of A are negated, and the resulting matrix A' is

$$A' = \begin{bmatrix} 3 & 0 & -2 \\ 0 & 9 & -2 \\ 4 & 2 & 7 \end{bmatrix}.$$

It is clear from our construction of A' that: (1) A' is at most rank-k, since $\mathrm{rank}(A') \leq \mathrm{rank}(A) \leq \dim(U) = k$; (2) A' is strictly diagonally dominant, i.e., $|A'_{ii}| > \sum_{j \neq i} |A'_{ij}|$ for every $i \in [k+1]$. However, these two are contradictory, since every strictly diagonally dominant matrix is full-rank by the Gershgorin circle theorem.

Bounding the number of dominant sets

Then, bounding the number of 1-dominant sets becomes a combinatorial problem: What is the largest possible size of a family $\mathcal{F} \subseteq \binom{[d]}{s}$ that contains no $k+1$ pairwise disjoint members? This problem has been studied since the work of Erdős [16], and the long-standing Erdős Matching Conjecture (EMC) states that when $d \geq s(k+1) - 1$, the maximum size is exactly

$$\max \left\{ \binom{d}{s} - \binom{d-k}{s}, \binom{s(k+1)-1}{s} \right\}.$$

Note that the above is the maximum between $|\mathcal{F}_1|$ and $|\mathcal{F}_2|$, where $\mathcal{F}_1 := \{S \in \binom{[d]}{s} : S \cap [k] \neq \emptyset\}$ and $\mathcal{F}_2 := \binom{[s(k+1)-1]}{s}$ are both families that contain no $k+1$ disjoint sets via simple pigeonhole arguments. A recent work of Frankl and Kupavskii [19] proved the EMC for all sufficiently large k and $d \geq \frac{5}{3}ks$.

While these proofs tend to be technical and sophisticated, if we are willing to lose a constant factor, a nearly-tight bound of $|\mathcal{F}| \leq O(ks/d) \cdot \binom{d}{s}$ is well-known and easy to prove via the following argument. Let $A_1, A_2, \ldots, A_{\lfloor d/s \rfloor}$ be pairwise disjoint members of $\binom{[d]}{s}$ chosen uniformly at random.[2] What would be the expected size of $\mathcal{F} \cap \{A_1, A_2, \ldots, A_{\lfloor d/s \rfloor}\}$? This number must be at most k; otherwise we would find $k+1$ disjoint sets in $\mathcal{F}$. Meanwhile, this expectation is exactly given by $|\mathcal{F}| \cdot \lfloor d/s \rfloor / \binom{d}{s}$, since each A_i coincides with each $S \in \mathcal{F}$ with probability $1/\binom{d}{s}$. The two observations together give $|\mathcal{F}| \leq k\binom{d}{s}/\lfloor d/s \rfloor = O(ks/d) \cdot \binom{d}{s}$. Furthermore, this bound is tight up to a constant factor, as the set family $\mathcal{F}_1$ defined earlier has size $\binom{d}{s} - \binom{d-k}{s} \geq \Omega(\min\{ks/d, 1\}) \cdot \binom{d}{s}$.

Handling weak dominance and the general t case

Recall that we made two simplifications: strict dominance and $t = 1$. Without the strict dominance, repeating the first half of the above argument would end up with a $(k+1) \times (k+1)$ matrix A' that is only weakly diagonally dominant, i.e., $|A'_{ii}| \geq \sum_{j \neq i} |A'_{ij}|$, which does not imply that A' is full rank.[3] Fortunately, we still have $\mathrm{rank}(A') = \Omega(k)$.[4] Thus, if we started with Ck independent 1-dominant sets for sufficiently large C, we would still obtain a contradiction to $\dim U = k$. The lower bound on $\mathrm{rank}(A')$ is proved by a probabilistic argument: We carefully sample a submatrix of A', so that the submatrix is strictly diagonally dominant, and the dimension only shrinks by a constant factor in expectation.

---

[2] More formally, we choose a permutation $\pi : [d] \to [d]$ uniformly at random, and let $A_i := \{\pi((i-1)s+1), \pi((i-1)s+2), \ldots, \pi(is)\}$.
[3] Consider the 2×2 all-one matrix.
[4] Technically, this also requires A' to have non-zero rows in addition to weak diagonal dominance.

To handle the general t case, we strengthen the combinatorial constraint on $\mathcal{F}$ into the following condition (Lemma 10, our main technical lemma):

The family of t-dominant sets does not contain $\Omega(k)$ sets that are "almost disjoint".

We will use two different notions of "almost disjointness": (1) the pairwise intersections are bounded by $O(t/k)$; (2) the sets become disjoint after removing $O(t)$ elements from each set. Note that the first definition implies the second, so, conversely, using the second notion imposes a stricter constraint on the family. Also, using either notion, the constraint gets stricter as t grows.

Lemma 10 is proved using an analogous (yet slightly more technical) argument to the above – assuming the existence of $\Omega(k)$ almost disjoint t-dominant sets, we construct a matrix using the witnesses of the dominance, such that the matrix has rank $\leq \dim U = k$ and satisfies a weaker diagonal dominance property. Finally, these two are shown to be contradictory via a probabilistic argument.

Using Lemma 10, we upper bound the number of t-dominant sets via two different approaches. The third bound in Theorem 2 follows from the first notion of almost disjointness, and an easy extension of the double counting argument presented above. The first bound in Theorem 2, on the other hand, relies on the finer-grained definition of disjointness and a quite different analysis based on Hall's marriage theorem.

2.2 Recovery of Subspace and the Entire Dataset

We briefly discuss how we recover the subspace in the setup of Problem 2, thereby reducing the problem to recovering the individual data points given the low-rank structure (i.e., Problem 1).

Let U be the column space of the Gaussian covariance Σ. For simplicity, we first assume $\mu = 0$ and that the first k coordinates uniquely determine a vector in U, i.e., there exist $c_{k+1}, c_{k+2}, \ldots, c_d \in \mathbb{R}^k$ such that for every $u \in U$ and $i \in \{k+1, k+2, \ldots, d\}$, $u_i = c_i^\top u|_{[k]}$. Then, the problem can be reduced to $d - k$ linear regression instances in k dimensions: Let Σ' be the $k \times k$ restriction of Σ to coordinates in $[k]$. For each clean data $x^{(j)} \in \mathbb{R}^d$, $\left(x^{(j)}|_{[k]}, x_i^{(j)} \right)$ can be viewed an instance drawn from $\mathcal{N}(0, \Sigma')$ and its label produced by the linear function $x \mapsto c_i^\top x$. Consequently, $\left(\tilde{x}^{(j)}|_{[k]}, \tilde{x}_i^{(j)} \right)$ is a potentially corrupted copy and the probability of corruption is at most $(k+1) \cdot s/d$, which is bounded by a small constant assuming $ks = O(d)$. Using the robust linear regression algorithm of [13], we can efficiently recover each c_i, and thus the whole subspace U.

In the general case, we can easily find a subset of k pivotal coordinates, since the algorithm of [13] also allows us to test whether including an additional coordinate introduces a linear dependence. Also note that the above procedure only recovers the column space U of the covariance, which might not contain the mean vector μ. To include μ into the subspace, it is sufficient to compute $\mu^\top v_1, \mu^\top v_2, \ldots, \mu^\top v_{d-k}$, where $\{v_i\}_{i \in [d-k]}$ is a basis of $U^\perp$. Note that for $x \sim \mathcal{N}(\mu, \Sigma)$, $x^\top v_i$ exactly gives $\mu^\top v_i$, though it might be altered if any of the coordinates in the support of v_i is corrupted. If we pick each v_i to be $(k+1)$-sparse (which is always possible), the fraction of incorrect values of $x^\top v_i$ is at most $(k+1) \cdot (s/d) \ll 1/2$. This allows us to recover $\mu^\top v_i$ by a simple majority vote.

Finally, we sketch the proof of Theorem 5: First, we find the subspace $U' = \mathrm{Span}(U \cup \{\mu\})$ as outlined above. Then, we estimate each μ_i up to an $O(Bs/d)$ additive error, by solving the one-dimensional mean estimation problem with corruption rate s/d and variance $\Sigma_{ii} \leq B^2$. We then clip the i-th coordinate of the data points so that they become $\tilde{O}(B)$-close to the

estimated mean; with high probability, no clean entries will be changed. Therefore, each clipped data point can be viewed as an instance of Problem 1 with known subspace U' and corruptions bounded by $\tilde{O}(B)$. Each data point can then be recovered up to a small ℓ_1 error using the BP method.

3   Related Work

Comparison to [Liu-Park-Rekatsinas-Tzamos, ICML'21]

Most closely related to our work is a recent paper of Liu, Park, Rekatsinas and Tzamos [33] on robust Gaussian mean estimation in a very similar setup. In their model, the clean data lie in the column space of a $d \times k$ matrix A (where $k \ll d$), and the adversary is allowed to corrupt a small fraction of coordinates in the dataset. An important difference is that we assume that the corruptions are randomly located, while the locations might be adversarially chosen in the model of [33]. In return, we make no assumptions on the subspace, whereas their estimation guarantee depends on a complexity measure of matrix A. Therefore, the results in the two models are qualitatively different and incomparable.

In more detail, the positive results of [33] rely on the assumption that the linear code defined by A has a large distance, which guarantees that the clean data can be decoded (information-theoretically) even with a small number of missing or corrupted coordinates. The mean estimation bounds in [33] depend on the quantity m_A defined as the smallest number m such that we can remove m rows from A to reduce its rank by 1. It is easy to show that this definition of m_A is equivalent to $\min_{z \in \mathbb{R}^k \setminus \{0\}} \|Az\|_0$, namely, the distance of the linear code defined by A.

In terms of the algorithm, we follow a similar approach to that of [33]: first denoising each data point using the low-rank structure, and then use the cleaned data to estimate the mean. While the assumptions of [33] makes the denoising information-theoretically easy, the decoding is NP-hard in general. In contrast, we analyze the natural ℓ_1 relaxation of the sparse recovery problem, which, via a non-trivial analysis, is shown to succeed with high probability over the randomness in the locations of the corrupted coordinates.

Other models of coordinate-level corruption

Chen, Caramanis and Mannor [9] considered a sparse linear regression setup, in which the weight vector has at most $k \ll d$ non-zero entries, and on each coordinate, the adversary may corrupt an ϵ fraction of the data entries (that are not necessarily randomly located). This corruption model is stronger than ours, since an ϵ fraction of the data points could be entirely corrupted. Their main result is an efficient algorithm that outputs a non-trivial estimate of the weight vector whenever the sample size is $\Omega(k \log d)$ and the corruption rate is $O\left(\frac{1}{\sqrt{k} \log d}\right)$.

Hu and Reingold [23] studied a robust mean estimation setting in which the data can be highly incomplete: the adversary is allowed to corrupt an ϵ fraction of the samples as well as to conceal all but a γ fraction of the entries on each coordinate. The concealed entries are chosen independently of the data. They gave efficient algorithms that recover the mean of the data distribution up to a dimension-independent ℓ_2 error of $\tilde{O}(\epsilon/\gamma)$ or $O(\sqrt{\epsilon/\gamma})$, depending on whether the distribution is assumed to be sub-Gaussian or just bounded-covariance.

Robust PCA

Robust PCA is the problem of decomposing a matrix into a low-rank component and a sparse component. In the seminal work of Candès, Li, Ma, and Wright [5], it was shown that under certain assumptions, such a decomposition is unique and can be computed efficiently using a simple convex program. Roughly speaking, the unique decomposition is possible when the low-rank component is far from being sparse (guaranteed by an appropriate incoherence assumption), and the sparse component is far from being low-rank (guaranteed by the assumption that the corruptions are randomly located). Subsequent work [22, 1, 43, 35] studied alternative formulations of the setup and optimization methods.

In the context of robust PCA, we may view the corrupted data in Problem 2 as an $n \times d$ matrix. Our main result (Theorem 5) guarantees the accurate recovery of the data matrix up to a small ℓ_1 error on each row. Compared to most of the work on robust PCA, our algorithm does not need any assumption on the low-rank subspace, at the cost of stronger distributional assumptions (i.e., the data are Gaussian), a more narrow parameter regime in which our results apply (i.e., the assumption that $ks = O(d)$), and an imperfect recovery guarantee (which is inevitable).

We remark that [1] is a notable exception: their assumptions on the low-rank component is milder than incoherence. When their result is specialized to robust PCA on a $d_1 \times d_2$ matrix, it requires each entry of the low-rank component to be bounded by $\alpha/\sqrt{d_1 d_2}$ in absolute value, where d_1 and d_2 are the dimensions of the matrix. In this scenario, their result implies the recovery of the low-rank part up to a squared Frobenius error of $O(\alpha^2 s'/(d_1 d_2))$, where s' is the number of corrupted entries. In the setup of our Problem 2, we have $s' = s \cdot n$ and $\alpha/\sqrt{d_1 d_2} = O(B\sqrt{\log(nd)})$ for $B = \sqrt{\max_{i \in [d]} \Sigma_{ii}}$.[5] Then, the result of [1] implies an $\tilde{O}(B^2 sn)$ squared Frobenius error, which is no better than using the corrupted data as the estimate. On the other hand, as noted in [1, Example 6], this bound cannot be further improved in their setup, since the locations of the sparse corruptions are allowed to be arbitrary.

Outlier-robust PCA

Another line of work on outlier-robust PCA [17, 42, 44, 27, 31, 11] studies a different and orthogonal model of robustness. In this setup, the goal is to identify the top eigenvector of an unknown distribution, given a corrupted dataset from that distribution. Compared to the robust PCA literature in the previous paragraph, this setting is further from our setting, since the corruptions are on the sample level, and the goal is to extract information of the underlying distribution, rather than recovering the dataset.

Sparse recovery

In sparse recovery, we aim to recover a k-sparse vector $x \in \mathbb{R}^d$ (also called the signal) given the value of $y = Ax$ for an $n \times d$ matrix A (also called the measurements). The Basis Pursuit program, minimizing $\|\hat{x}\|_1$ subject to $A\hat{x} = y$, is a natural convex relaxation for the problem. Unlike the setup of our Problem 1, the linear measurement A is often under the control of the recovery algorithm, and the goal is to design A such that efficient recovery is possible; in contrast, the linear measurements (induced by the unknown subspace U) are arbitrary in

---

[5] We may first estimate each coordinate up to an $O(Bs/d)$ error, and subtract this estimate from the data. After the shifting, the clean data are bounded by $O(B\sqrt{\log(nd)})$ entrywise.

Problem 1. The sparse recovery problem is also closely related to compressed sensing [6, 14] as well as data stream algorithms. A different formulation has been studied in the TCS literature: even when x is not k-sparse, we are still required to output a k-sparse $\hat{x}$ such that $\|\hat{x} - x\|_p$ is close to $\min_{\|x'\|_0 \leq k} \|x' - x\|_p$. Nearly tight performance guarantees have been obtained for various different metrics of performance guarantee [26, 2, 37, 24], and also in a related setup where the linear measurements can be chosen adaptively [25, 38, 29].

Robust subspace recovery

Robust subspace recovery is the problem of identifying a subspace that contains the *inliers* (i.e., clean data) in the presence of outliers that might be outside this subspace. [21] gave an algorithm that is provably correct and efficient as long as: (1) The fraction of inliers is at least k/d, where k and d are the dimensions of the subspace and the ambient space, respectively; (2) The inliers and outliers together satisfy a non-degeneracy assumption. [34] studies the case where only the inliers are assumed to be in general position and the outliers can be arbitrary. Recent works of [39, 3] gave algorithms in the list-decodable learning setting based on the sum-of-squares method, under additional distributional assumptions on the inliers. See [32] for a detailed survey of this literature.

Extremal combinatorics

As outlined in Section 2, our analysis of the BP method involves upper bounding the size of a set family subject to certain combinatorial constraints (e.g., on the pairwise intersections between the sets in the family). This is the focus of extremal combinatorics. In particular, our proof of Lemma 12 is based on techniques in this literature. For many problems in extremal combinatorics, the goal is to either find the exact maximum, or at least obtain asymptotically tight bounds (i.e., tight up to a $1 + o(1)$ factor). In some cases, the set families that achieve this maximum size have also been characterized. In contrast, for the purpose of this work, we are more interested in bounds that might be off by a constant factor, but hold for a wider range of parameters.

Most closely related to our approach is the Erdős Matching Conjecture regarding families of size-s subsets of $[d]$ that do not contain $k + 1$ pairwise disjoint sets. [16] conjectured that the maximum size is given by the larger between $\binom{d}{s} - \binom{d-k}{s}$ and $\binom{s(k+1)-1}{s}$, for all $d \geq s(k+1) - 1$. It was also shown in [16] that the first term gives the correct answer for fixed k, s and sufficiently large $d \geq d_0(k, s)$. This result was strengthened by subsequent work, and Frankl and Kupavskii [19] recently relaxed the assumption to $d \geq \frac{5}{3}ks$ for all sufficiently large k; we refer the readers to [19] for a more detailed discussion on the progress towards proving EMC. Another notable special case is the $k = 1$ case, in which the family is not allowed to contain any two disjoint sets. A more general version of this case, now known as the Erdős-Ko-Rado Theorem, was proved in [15]; see [30] for a simplified proof and [10] for a survey of several generalizations and analogues of this theorem.

Part of our proof (Lemma 13) also requires constructing a large *packing*, i.e., a subset of $\binom{[d]}{s}$ such that the size of any pairwise intersection among the family is at most $\Delta - 1$.[6] It is easy to see that the size of any packing is at most $\frac{\binom{d}{\Delta}}{\binom{s}{\Delta}}$: every member in the packing contains $\binom{s}{\Delta}$ subsets of size Δ, and no two members may share the same size-Δ subset. A Theorem of Rödl [40] shows that this bound is asymptotically tight: for fixed s, Δ and as

---

[6] This is also called a "combinatorial design" in the TCS literature.

$d \to +\infty$, the largest packing has size $(1 - o(1)) \cdot \frac{\binom{d}{\Delta}}{\binom{s}{\Delta}}$. Unfortunately, this theorem cannot be directly applied, since it does not hold uniformly for all values of (d, s, Δ) (even up to a constant factor). Instead, we use a well-known construction based on polynomials (which, e.g., was used by Nisan [36] for constructing pseudorandom generators) that nearly achieves this limit in the $s = O(\sqrt{d})$ regime.

4   Proof of the Sparse Recovery Guarantee

In this section, we prove Theorem 2, which states that every optimum x^* of the Basis Pursuit (BP) program below is close to the clean data x with high probability.

$$\min_{\hat{x}} \quad \|\hat{x} - \tilde{x}\|_1$$
$$\text{subject to} \quad \hat{x} \in U$$

4.1   A Necessary Condition for Large Error

We start by stating a simple necessary condition for BP to incur a large error. Recall that in Problem 1, U is the subspace in which x lies and $S \subseteq [d]$ is the set of corrupted coordinates.

▶ **Lemma 9.** *In the setup of Problem 1 with $B = 1$, if BP incurs an ℓ_1 error of $t > 0$, S is t-dominant with respect to U in the sense that there exists $u \in U$ such that $\|u\|_1 = 1$ and*

$$\sum_{i \in S} |u_i| \cdot \mathbb{1}\left[|u_i| \leq \frac{1}{t}\right] \geq \frac{1}{2}.$$

Note that when $t \leq 1$, the inequality in Lemma 9 reduces to $\|u|_S\|_1 \geq 1/2$, which is exactly the negation of the restricted nullspace property (Definition 7).

Proof. We first note that it is sufficient to analyze the $x = 0$ case, since both the problem setting and the BP method are invariant under a translation within U. Formally, let $\tilde{x}$ be the corrupted version of x, and x^* be an optimum of BP (Program (1)). Then, for the alternative clean data $x' = 0$, $(\tilde{x})' := \tilde{x} - x$ is a valid corrupted version of x' (in the setup of Problem 1), and $(x^*)' := x^* - x$ would be an optimum of BP. The error would still be $\|(x^*)' - x'\|_1 = \|x^* - x\|_1 = t$.

Let $\tilde{x}$ be a corrupted copy of the clean data $x = 0$. Let $x^* \in \operatorname{argmin}_{\hat{x} \in U} \|\hat{x} - \tilde{x}\|_1$ be an optimum with $\|x^*\|_1 = t$ and $u := \frac{x^*}{\|x^*\|_1}$. Clearly, $u \in U$ and $\|u\|_1 = 1$. Furthermore, since $tu = x^* \in \operatorname{argmin}_{\hat{x} \in U} \|\hat{x} - \tilde{x}\|_1$, we have $tu \in \operatorname{argmin}_{\hat{x} \in \operatorname{Span}(\{u\})} \|\hat{x} - \tilde{x}\|_1$, or equivalently,

$$t \in \operatorname*{argmin}_{\alpha \in \mathbb{R}} \|\alpha u - \tilde{x}\|_1.$$

The objective function above can be expanded as

$$\|\alpha u - \tilde{x}\|_1 = \sum_{i=1}^{d} |u_i \alpha - \tilde{x}_i| = \sum_{i \in [d]: u_i \neq 0} |u_i| \cdot \left|\alpha - \frac{\tilde{x}_i}{u_i}\right| + \sum_{i \in [d]: u_i = 0} |\tilde{x}_i|.$$

This is equivalent to finding the weighted median of d points on a line, where the i-th point is at location $\frac{\tilde{x}_i}{u_i}$ and has a weight of $|u_i|$. Therefore, in order for t be a minimum, the weight in the region $[t, +\infty)$ must be at least half of the total weight (which is $\|u\|_1 = 1$), i.e.,

$$\sum_{i=1}^{d} |u_i| \cdot \mathbb{1}\left[\frac{\tilde{x}_i}{u_i} \geq t\right] \geq \frac{1}{2}.$$

Recall that in Problem 1, $\tilde{x}$ is required to satisfy: (1) $|\tilde{x}_i| \leq B = 1$; (2) $\tilde{x}_i = 0$ for $i \notin S$. Thus, the left-hand side above is maximized when we pick $\tilde{x}_i = 0$ for $i \notin S$ and $\tilde{x}_i = \text{sgn}(u_i)$ for every $i \in S$. This gives the desired inequality $\sum_{i \in S} |u_i| \cdot \mathbb{1}\left[|u_i| \leq \frac{1}{t}\right] \geq \frac{1}{2}$.   ◄

4.2   A Combinatorial Constraint on Dominant Subsets

It remains to show that, for large t and *any* low-dimensional subspace U, over the randomness in the corruption set S, S is unlikely to be t-dominant as defined in Lemma 9. Formally, we fix the k-dimensional subspace $U \subseteq \mathbb{R}^d$, size $s \in [d]$ and the error parameter $t > 0$. Define $\mathcal{F}_t^{(s)}$ as the family of size-s subsets of $[d]$ that are t-dominant, i.e.,

$$\mathcal{F}_t^{(s)} := \left\{ S \in \binom{[d]}{s} : \exists u \in U \text{ such that } \|u\|_1 = 1, \sum_{i \in S} |u_i| \cdot \mathbb{1}\left[|u_i| \leq 1/t\right] \geq \frac{1}{2} \right\}.$$

To upper bound the size of $\mathcal{F}_t^{(s)}$, we give the following combinatorial property of this family.

▶ **Lemma 10.** *Let* $\alpha \in \left(0, \frac{1}{2}\right)$, $k' := \left\lfloor \frac{4(1-\alpha)k}{\alpha} \right\rfloor + 1$, $t' := \lfloor (1/2 - \alpha)t \rfloor$ *and* $\Delta := \left\lfloor \frac{t'}{k'-1} \right\rfloor$. *For any* k' *sets* $S_1, S_2, \ldots, S_{k'} \in \mathcal{F}_t^{(s)}$:
   ■ *There are no pairwise disjoint sets* $T_1, T_2, \ldots, T_{k'}$ *such that* $T_i \subseteq S_i$ *and* $|S_i \setminus T_i| \leq t'$ *for every* $i \in [k']$.
   ■ *In particular, there must be a pair* S_i *and* S_j *such that* $|S_i \cap S_j| > \Delta$ *and* $i \neq j$.

Concretely, by setting $\alpha = 1/4$, we get $k' = 12k + 1$, $t' = \lfloor t/4 \rfloor$ and $\Delta = \lfloor t/(48k) \rfloor$. In general, for every fixed α, the first part of Lemma 10 says that we cannot find $\Omega(k)$ sets in $\mathcal{F}_t^{(s)}$ that become disjoint after removing $O(t)$ elements from each, while the "in particular" part says that $\mathcal{F}_t^{(s)}$ cannot contain $\Omega(k)$ sets such that every pairwise intersection is of size $O(t/k)$.

Proof. Suppose towards a contradiction that $S_1, \ldots, S_{k'}$ are k' sets in $\mathcal{F}_t^{(s)}$ that violate either of the two conditions. We will derive a contradiction in the following three steps: First, we show that violating the second condition implies a violation of the first. Then, we use the fact that the sets are in $\mathcal{F}_t^{(s)}$, together with the violation of the first condition, to obtain a matrix with a diagonal dominance property. Such a matrix is constructed so that its rank is at most the dimension of U. Finally, we show that the diagonal dominance implies a lower bound on the rank of the matrix, which contradicts the assumption that U is low-dimensional.

Reduce to the first condition.   Suppose that the second condition is violated, i.e., there are sets $S_1, \ldots, S_{k'} \in \mathcal{F}_t^{(s)}$ with small ($\leq \Delta$) pairwise intersections. Then, we define

$$T_i := S_i \setminus \left(\bigcup_{j \in [k'] \setminus \{i\}} S_j \right).$$

Clearly, $T_1, T_2, \ldots, T_{k'}$ are pairwise disjoint and each T_i is a subset of S_i. Furthermore, $|S_i \setminus T_i|$ is at most $\sum_{j \in [k'] \setminus \{i\}} |S_i \cap S_j| \leq (k'-1)\Delta \leq t'$. This shows that the first condition is also violated.

Construct a diagonally dominant matrix.   For each $i \in [k']$, let $u^{(i)} \in U$ be a witness of $S_i \in \mathcal{F}_t^{(s)}$, i.e., $\|u^{(i)}\|_1 = 1$ and

$$\sum_{j \in S_i} |u_j^{(i)}| \cdot \mathbb{1}\left[|u_j^{(i)}| \leq 1/t\right] \geq \frac{1}{2}. \tag{2}$$

It follows that

$$
\sum_{j \in T_i} |u_j^{(i)}| \geq \sum_{j \in T_i} |u_j^{(i)}| \cdot \mathbb{1}\left[|u_j^{(i)}| \leq 1/t\right]
$$

$$
= \sum_{j \in S_i} |u_j^{(i)}| \cdot \mathbb{1}\left[|u_j^{(i)}| \leq 1/t\right] - \sum_{j \in S_i \setminus T_i} |u_j^{(i)}| \cdot \mathbb{1}\left[|u_j^{(i)}| \leq 1/t\right] \qquad (T_i \subseteq S_i)
$$

$$
\geq \frac{1}{2} - |S_i \setminus T_i| \cdot \frac{1}{t} \qquad \text{(Equation (2))}
$$

$$
\geq \frac{1}{2} - \left(\frac{1}{2} - \alpha\right) = \alpha. \qquad (|S_i \setminus T_i| \leq (1/2 - \alpha)t)
$$

In words, an α-fraction of the 1-norm of $u^{(i)}$ is contributed by the coordinates in T_i.

Now, we construct a $k' \times k'$ matrix A using the vectors $u^{(1)}, \ldots, u^{(k')}$, and then argue that the rank of A is at most k. Roughly speaking, each entry A_{ij} is obtained by aggregating the entries $\left\{u_l^{(i)} : l \in T_j\right\}$ with some of the entries negated, so that the diagonal entries of A are maximized. Formally, we pick $s \in \{\pm 1\}^d$ such that:

$$
s_j = \begin{cases} +1, & j \in T_i \text{ and } u_j^{(i)} \geq 0 \text{ for some } i, \\ -1, & j \in T_i \text{ and } u_j^{(i)} < 0 \text{ for some } i, \\ \text{arbitrary}, & j \notin \bigcup_{i \in [k']} T_i. \end{cases}
$$

Then, we let $A_{ij} := \sum_{l \in T_j} s_j u_j^{(i)}$.

It is easy to verify that A is well-defined, and the following hold for every $i \in [k']$:

- $A_{ii} = \sum_{j \in T_i} \left|u_j^{(i)}\right| \geq \alpha.$
- $\sum_{j=1}^{k'} |A_{ij}| \leq \sum_{j \in [k']} \sum_{l \in T_j} |u_l^{(i)}| \leq \left\|u^{(i)}\right\|_1 = 1.$

Furthermore, A can be obtained from the $k' \times d$ matrix

$$
\begin{bmatrix} u^{(1)} & u^{(2)} & \cdots & u^{(k')} \end{bmatrix}^\top
$$

via a sequence of elementary column operations (specifically, swapping two columns and adding a column or its negation to another column) and column deletions. Thus, the rank of A is at most $\dim U = k$.

Lower bound the rank via a probabilistic argument. We scale the rows of A, so that the absolute values of the entries in each row sum up to 1. Clearly, in the resulting matrix, we still have $A_{ii} \geq \alpha$ for each $i \in [k']$.

We consider a randomly chosen set $J \subseteq [k']$, such that each element is included in J with probability $p := \frac{\alpha}{2 \cdot (1-\alpha)}$ independently. We say that index $i \in [k']$ is *dominant* if: (1) $i \in J$; (2) $\sum_{j \in J \setminus \{i\}} |A_{ij}| < \alpha$. By our choice of J, the two events are independent. The probability of the former is p, while the probability of the latter is, by Markov's inequality, at least

$$
1 - \frac{\mathbb{E}_J\left[\sum_{j \in J \setminus \{i\}} |A_{ij}|\right]}{\alpha} = 1 - \frac{p \cdot \sum_{j \in [k'] \setminus \{i\}} |A_{ij}|}{\alpha} \geq 1 - \frac{p \cdot (1-\alpha)}{\alpha} = \frac{1}{2}.
$$

Therefore, each index i is dominant with probability at least $p \cdot \frac{1}{2} = \frac{\alpha}{4 \cdot (1-\alpha)}$.

By an averaging argument, there exists a deterministic set $J \subseteq [k']$ such that the number of dominant indices is at least $\left\lceil \frac{\alpha}{4 \cdot (1-\alpha)} \cdot k' \right\rceil > k$. Let $J' \subseteq J$ be the subset of dominant indices. Then, the $|J'| \times |J'|$ submatrix of A induced by J' is diagonally dominant, and thus full-rank by the Gershgorin circle theorem. This implies $\text{rank}(A) \geq |J'| > k$, which contradicts our ealier conclusion that $\text{rank}(A) \leq k$. ◀

4.3 Bounding the Number of Dominant Subsets

Next, we upper bound $|\mathcal{F}_t^{(s)}|$ using Lemma 10. We use the following definition of packings.

▶ **Definition 11.** *A (d, s, Δ)-packing is a subset of $\binom{[d]}{s}$ such that the intersection of any two different sets has size at most $\Delta - 1$.*

With Definition 11, the "in particular" part of Lemma 10 simply states that "$\mathcal{F}_t^{(s)}$ does not contain a $(d, s, \Delta + 1)$-packing of size k'".

Let $m(d, s, \Delta)$ denote the size of the maximum (d, s, Δ)-packing. The following lemma gives an upper bound on $|\mathcal{F}_t^{(s)}|$ in terms of the packing size.

▶ **Lemma 12.** *For k' and Δ chosen as in Lemma 10,*

$$|\mathcal{F}_t^{(s)}| \leq \frac{k' - 1}{m(d, s, \Delta + 1)} \cdot \binom{d}{s}.$$

Lemma 12 implies that whenever $m(d, s, \Delta + 1) \gg k'$, $\mathcal{F}_t^{(s)}$ only contains a negligible fraction of $\binom{[d]}{s}$. The lemma is proved by studying the intersection between $\mathcal{F}_t^{(s)}$ and a $(d, s, \Delta + 1)$-packing after a random permutation over the elements. This technique appears in a proof of the Erdös-Ko-Rado theorem [15] due to Katona [30]; also see [28, Section 7.2] for an exhibition.

Proof. Let $M := m(d, s, \Delta + 1)$ and $\{A_1, A_2, \ldots, A_M\}$ be a $(d, s, \Delta + 1)$-packing. Let π be a permutation of $[d]$ chosen uniformly at random. Note that for each $i \in [M]$, $\pi(A_i) := \{\pi(e) : e \in A_i\}$ is still a size-s subset of $[d]$. We examine the quantity

$$\mathbb{E}_\pi \left[\sum_{S \in \mathcal{F}_t^{(s)}} \sum_{i=1}^{M} \mathbb{1} \left[S = \pi(A_i) \right] \right].$$

For each fixed π, $\pi(A_1), \pi(A_2) \ldots, \pi(A_M)$ is still a $(d, s, \Delta + 1)$-packing, and the double summation inside the expectation is simply the size of $\mathcal{F}_t^{(s)} \cap \{\pi(A_1), \pi(A_2) \ldots, \pi(A_M)\}$. Such an intersection must have size at most $k' - 1$ in light of Lemma 10. On the other hand, the quantity is equal to, by the linearity of expectation,

$$\sum_{S \in \mathcal{F}_t^{(s)}} \sum_{i=1}^{M} \Pr_\pi \left[S = \pi(A_i) \right] = \frac{|\mathcal{F}_t^{(s)}| \cdot M}{\binom{d}{s}}.$$

Therefore, we have $|\mathcal{F}_t^{(s)}| \leq \frac{k'-1}{M} \cdot \binom{d}{s}$.   ◀

Finally, the following lemma states two simple lower bounds on the packing number $m(d, s, \Delta)$.

▶ **Lemma 13.** *Suppose that q is a prime power, $1 \leq \Delta \leq s \leq q$ and $d \geq sq$. We have*

$$m(d, s, \Delta) \geq q^\Delta.$$

Furthermore, for any d and s, we have $m(d, s, 1) \geq \lfloor d/s \rfloor$.

Proof. Let $\mathbb{F}_q$ be the finite field of size q. For each $a = (a_0, a_1, \ldots, a_{\Delta-1}) \in \mathbb{F}_q^\Delta$, we consider the polynomial $p_a(x) := a_0 + a_1 x + \cdots + a_{\Delta-1} x^{\Delta-1}$ over $\mathbb{F}_q$. Since $s \leq q$, we can find s distinct elements $x_1, x_2, \ldots, x_s \in \mathbb{F}_q$. We then define the set associated with a as

$$S_a := \{(1, p_a(x_1)), (2, p_a(x_2)), \ldots, (s, p_a(x_s))\}.$$

Clearly, S_a is a size-s subset of $[s] \times \mathbb{F}_q$. Furthermore, for $a \neq a'$, since $p_a(x) - p_{a'}(x)$ is a non-zero polynomial of degree $< \Delta$, $p_a(x)$ and $p_{a'}(x)$ agree on at most $\Delta - 1$ points. This implies $|S_a \cap S_{a'}| \leq \Delta - 1$. Since $d \geq sq$, via an arbitrary injection from $[s] \times \mathbb{F}_q$ to $[d]$, $\{S_a : a \in \mathbb{F}_q^{\Delta}\}$ gives a (d, s, Δ)-packing of size q^{Δ}.

Finally, the "furthermore" part follows from the $(d, s, 1)$-packing given by $\{1, 2, \ldots, s\}$, $\{s + 1, s + 2, \ldots, 2s\}$, ..., $\{s(\lfloor d/s \rfloor - 1) + 1, \ldots, s\lfloor d/s \rfloor\}$. ◀

Combining Lemmas 9, 10, 12 and 13 immediately gives part of Theorem 2.

Proof of Theorem 2 (the second and third bounds). By Lemma 9 and the definition of $\mathcal{F}_t^{(s)}$, the probability $\Pr\left[\|x^* - x\|_1 \geq t\right]$ is at most by $|\mathcal{F}_t^{(s)}|/\binom{d}{s}$, which is, by Lemma 12, at most $(k' - 1)/m(d, s, \Delta + 1)$. Choosing $\alpha = 1/4$ in Lemma 10 gives $k' = 12k + 1$, $\Delta = \lfloor t/(48k) \rfloor$ and

$$\Pr\left[\|x^* - x\|_1 \geq t\right] \leq \frac{12k}{m(d, s, \lfloor t/(48k) \rfloor + 1)}.$$

Now we apply Lemma 13 by choosing q as the unique power of 2 in $\left(\frac{d}{2s}, \frac{d}{s}\right]$. Since $s \leq \sqrt{d/2}$, we have $s \leq \frac{d}{2s} < q$ and $d \geq sq$, i.e., the preconditions of Lemma 13 are satisfied, and we get

$$m(d, s, \lfloor t/(48k) \rfloor + 1) \geq q^{\lfloor t/(48k) \rfloor + 1} \geq \left(\frac{d}{2s}\right)^{\lfloor t/(48k) \rfloor + 1}.$$

This gives the third bound $\Pr\left[\|x^* - x\|_1 \geq t\right] \leq 12k \cdot \left(\frac{2s}{d}\right)^{\lfloor t/(48k) \rfloor + 1}$.

To prove the second bound, we take $t_0 = \min\{t, 24k\}$, and the above argument gives

$$\Pr\left[\|x^* - x\|_1 \geq t\right] \leq \Pr\left[\|x^* - x\|_1 \geq t_0\right] \leq \frac{12k}{m(d, s, 1)}.$$

By the "furthermore" part of Lemma 13, $m(d, s, 1) \geq \lfloor d/s \rfloor \geq d/(2s)$. This bounds the last term above by $24ks/d$. ◀

4.4   Applying the Stricter Constraint

To prove the first bound in Theorem 2, we need to apply the first part of Lemma 10 (instead of the "in particular" part) to obtain the following alternative upper bound on $|\mathcal{F}_t^{(s)}|$.

▶ **Lemma 14.** *For k' and t' chosen as in Lemma 10,*

$$|\mathcal{F}_t^{(s)}| \leq \binom{s(k' - 1)}{t' + 1} \cdot \binom{d - t' - 1}{s - t' - 1}.$$

To prove Lemma 14, we define the following notion of a perfect matching and state a characterization of its existence from Hall's theorem.

▶ **Definition 15** (Perfect s-matching.). *A collection of n sets $S_1, S_2, \ldots, S_n$ can be perfectly s-matched if there are pairwise disjoint sets $T_1 \in \binom{S_1}{s}$, $T_2 \in \binom{S_2}{s}$, ..., $T_n \in \binom{S_n}{s}$.*

With this definition, the main part of Lemma 10 can be re-written as "no k' sets in $\mathcal{F}_t^{(s)}$ can be perfectly $(s - t')$-matched".

We can view the above definition in terms of a bipartite graph (X, Y, E), in which the vertex sets are $X = [n]$ and $Y = \bigcup_{i \in [n]} S_i$, while $(x, y) \in X \times Y$ is present in the edge set E if and only if $y \in S_x$. The s-matching in the above definition then corresponds to a subset of E such that each X-vertex has exactly s incident edges, while each Y-vertex has at most one incident edge.

The following immediate corollary of Hall's marriage theorem characterizes the existence of perfect s-matchings.

▶ **Lemma 16** (Hall's marriage theorem [20]). *$S_1, S_2, \ldots, S_n$ can be perfectly s-matched if and only if for every $J \subseteq [n]$,*

$$\left| \bigcup_{i \in J} S_i \right| \geq |J| \cdot s.$$

Proof. The "only if" direction is easy: If a perfect s-matching is witnessed by $T_1, T_2, \ldots, T_n$, it holds for every $J \subseteq [n]$ that

$$\left| \bigcup_{i \in J} S_i \right| \geq \left| \bigcup_{i \in J} T_i \right| = \sum_{i \in J} |T_i| = |J| \cdot s.$$

For the "if" direction, suppose that the inequality holds for every $J \subseteq [n]$. We construct another bipartite graph $G = (X, Y, E)$ by representing each of the n sets with s vertices. Formally, $X = [n] \times [s]$, $Y = \bigcup_{i=1}^n S_i$ and

$$E = \{((i, j), y) \in X \times Y : y \in S_i\}.$$

Let $N(x) := \{y \in Y : (x, y) \in E\}$ denote the neighbours of $x \in X$.

We claim that $|\bigcup_{x \in J} N(x)| \geq |J|$ holds for every $J \subseteq X$. Assuming this claim, we are done: Hall's marriage theorem shows that G has a matching of size $|X|$. In other words, there exists an injection $f : [n] \times [s] \to \bigcup_{i=1}^n S_i$ such that $f(i, j) \in S_i$. Then, picking $T_i := \{f(i, j) : j \in [s]\}$ shows that $S_1, S_2, \ldots, S_n$ can be perfectly s-matched.

To verify the claim, let $J' := \{i \in [n] : \exists j \in [s], (i, j) \in J\}$ be the projection of J onto the first coordinate. Clearly, $|J| \leq |J'| \cdot s$. By our construction of G,

$$\bigcup_{x \in J} N(x) = \bigcup_{(i,j) \in J} S_i = \bigcup_{i \in J'} S_i.$$

Then, our assumption implies that the cardinality of the last set above is at least $|J'| \cdot s \geq |J|$. This establishes the inequality $|\bigcup_{x \in J} N(x)| \geq |J|$ and completes the proof. ◄

Now we are ready to prove Lemma 14.

Proof of Lemma 14. Let k^* be the largest integer such that there exist $S_1, S_2, \ldots, S_{k^*} \in \mathcal{F}_t^{(s)}$ that can be perfectly $(s - t')$-matched. By Lemma 10, k^* is at most $k' - 1$. Let $U := S_1 \cup S_2 \cup \cdots \cup S_{k^*}$ be the union of the k^* sets. We have $|U| \leq sk^* \leq s(k' - 1)$.

In the rest of the proof, we will show that every $T \in \mathcal{F}_t^{(s)}$ must intersect U at $\geq t' + 1$ elements. The lemma would then immediately follow, since every member of $\mathcal{F}_t^{(s)}$ can be specified by choosing a size-$(t'+1)$ subset of U and then choosing $s - t' - 1$ from the remaining $d - t' - 1$ elements in $[d]$. The number of ways of doing this is at most

$$\binom{|U|}{t' + 1} \cdot \binom{d - t' - 1}{s - t' - 1} \leq \binom{s(k' - 1)}{t' + 1} \cdot \binom{d - t' - 1}{s - t' - 1}$$

as desired.

Suppose towards a contradiction that some $T \in \mathcal{F}_t^{(s)}$ satisfies $|T \cap U| \leq t'$. We write $S_{k^*+1} := T$ and show that S_1 through S_{k^*+1} can also be perfectly $(s - t')$-matching, which violates the maximality of k^*. In light of Lemma 16, it suffices to prove $\left| \bigcup_{i \in J} S_i \right| \geq (s - t')|J|$ for every $J \subseteq [k^* + 1]$. If $k^* + 1 \notin J$, the assumption that $S_1, \ldots, S_{k^*}$ can be perfectly $(s - t')$-matched together with Lemma 16 gives $\left| \bigcup_{i \in J} S_i \right| \geq (s - t')|J|$. If $k^* + 1 \in J$, the same argument gives

$$\left| \bigcup_{i \in J \setminus \{k^*+1\}} S_i \right| \geq (s - t')(|J| - 1).$$

Furthermore, we have

$$\left| S_{k^*+1} \cap \left(\bigcup_{i \in J \setminus \{k^*+1\}} S_i \right) \right| \leq \left| S_{k^*+1} \cap \left(\bigcup_{i \in [k^*]} S_i \right) \right| = |T \cap U| \leq t'.$$

The two inequalities above together give

$$\left| \bigcup_{i \in J} S_i \right| = \left| \bigcup_{i \in J \setminus \{k^*+1\}} S_i \right| + |S_{k^*+1}| - \left| S_{k^*+1} \cap \left(\bigcup_{i \in J \setminus \{k^*+1\}} S_i \right) \right|$$

$$\geq (s - t')(|J| - 1) + s - t' = (s - t')|J|.$$

Therefore, applying Lemma 16 again shows that $S_1, \ldots, S_{k^*+1}$ can be perfectly $(s - t')$-matched, which leads to a contradiction.   ◄

Finally, we prove the first bound in Theorem 2.

Proof of Theorem 2 (the first bound). Lemmas 9 and 14 together imply that for some k' and t' chosen as in Lemma 10:

$$\Pr\left[\|x^* - x\|_1 \geq t\right] \leq \frac{|\mathcal{F}_t^{(s)}|}{\binom{d}{s}} \tag{Lemma 9}$$

$$\leq \frac{\binom{s(k'-1)}{t'+1} \cdot \binom{d-t'-1}{s-t'-1}}{\binom{d}{s}} \tag{Lemma 14}$$

$$\leq \frac{[s(k'-1)]^{t'+1}}{(t'+1)!} \cdot \frac{s}{d} \cdot \frac{s-1}{d-1} \cdots \frac{s-t'}{d-t'}$$

$$\left(\binom{n}{m} \leq n^m/m! \text{ and } \binom{n}{m} = \frac{n}{m}\binom{n-1}{m-1} \right)$$

$$\leq \frac{1}{(t'+1)!} \cdot \left[\frac{s^2(k'-1)}{d} \right]^{t'+1}.$$

The bound then follow from choosing $\alpha = 1/4$, $k' = 12k + 1$ and $t' = \lfloor t/4 \rfloor$ in Lemma 10.   ◄

5   Discussion and Open Problems

In this section, we highlight a concrete open problem, the solutions of which could further strengthen the recovery guarantees presented in this work. We also discuss several directions of further exploration on robust algorithms under coordinate-level corruptions.

5.1   Optimal Error under Weaker Assumptions?

Recall that in Theorem 5, the nearly-optimal error bound of $\tilde{O}(Bks/d)$ only holds in the $ks^2 = O(d)$ regime. In comparison, the rest of the proof would go through under the weaker assumption of $ks = O(d)$. We discuss a concrete approach to closing this gap.

This extra factor of s can be traced back to Lemma 14, which we restate below (after replacing k' and t' with k and t for brevity). Recall the definition of a perfect s-matching from Definition 15.

▶ **Lemma 14** (rephrased). *If $\mathcal{F} \subseteq \binom{[d]}{s}$ contains no k sets that can be perfectly $(s-t)$-matched,*

$$|\mathcal{F}| \leq \binom{s(k-1)}{t+1} \cdot \binom{d-t-1}{s-t-1}.$$

We conjecture that maximum possible size of $\mathcal{F}$ is achieved by one of a few simple constructions.

▶ **Conjecture 17.** *If $\mathcal{F} \subseteq \binom{[d]}{s}$ contains no k sets that can be perfectly $(s-t)$-matched,*

$$|\mathcal{F}| \leq \max_{i \in [s-t]} |\mathcal{F}_i|,$$

where $\mathcal{F}_i := \left\{ S \in \binom{[d]}{s} : |S \cap [ik-1]| \geq t + i \right\}$.

We first note that each $\mathcal{F}_i$ defined in Conjecture 17 is a valid choice of $\mathcal{F}$ – no k sets $S_1, \ldots, S_k \in \mathcal{F}_i$ can be perfectly $(s-t)$-matched. This is because for any $T_j \subseteq S_j$ of size $s - t$, $|T_j \cap [ik-1]|$ is at least $|S_j \cap [ik-1]| - t \geq i$. By a pigeonhole argument, $T_1, T_2, \ldots, T_k$ cannot be disjoint. Thus, if $\max |\mathcal{F}_i|$ is indeed a valid upper bound, it cannot be improved any further. The idea behind the construction of $\mathcal{F}_i$ is to restrict a certain number of degrees of freedom (namely, $t + i$) to a small subset (namely, $[ik-1]$), and then allow the remaining elements to be freely chosen. We remark that constructions that resemble the above have been shown or conjectured to be optimal for many natural problems in extremal combinatorics, including those in the Erdős-Ko-Rado theorem and the Erdős Matching Conjecture. We have also verified Conjecture 17 for $d \leq 8$, $s \in [d]$, $k \in \{2, 3, \ldots, d\}$ and $t \in \{0, 1, \ldots, s-1\}$, either manually or via an exhaustive search.

The upper bound in Conjecture 17 can be relaxed to the following closed-form expression:

$$|\mathcal{F}| \leq \left(\frac{eks}{d} \right)^{t+1} \cdot \binom{d}{s}.$$

Then, repeating the proof of Theorem 2 and Corollary 3 would give an $\tilde{O}(ks/d)$ bound whenever $ks \leq c_0 d$ for some sufficiently small constant $c_0 > 0$. Details can be found in the full version.

5.2 The $ks \gg d$ Regime?

Our results for recovering the subspace and the individual data points crucially rely on the assumption that $ks = O(d)$. For subspace recovery, this assumption guarantees that when restricting the problem to any $O(k)$ coordinates, a majority of data points will be entirely clean, which allows the known robust estimators (under sample-level corruption) to achieve an accurate recovery.

Even when the subspace is given, when $ks \gg d$, the expected error bound in Corollary 3 reduces to $\tilde{O}(Bs \cdot (ks/d)^2)$, which is looser than the trivial upper bound of $O(Bs)$.[7] This seems to be an inherent limitation of the current approach. When $ks \gg d$, our constraint on $\mathcal{F}_t^{(s)}$ in Lemma 10 is vacuous even when $t = s$: Even if $\mathcal{F}_t^{(s)} = \binom{[d]}{s}$, we still cannot find $k' = \Theta(k)$ sets that can be perfectly $(s-t')$-matched for $t' \leq t/2 = s/2$, which is easily seen by the pigeonhole principle and that $k'(s-t') = \Omega(ks) \gg d$. Consequently, the current approach would fail to control the probability that BP incurs an $\Omega(s)$ error. On the other hand, natural choices for the subspace U do not seem to give counterexamples in which $|\mathcal{F}_t^{(s)}|$ is close to $\binom{d}{s}$, even if $ks \gg d$. Therefore, obtaining non-trivial guarantees in the $ks \gg d$ regime *might* still be possible based on the generalized version of RNP (Lemma 9), though controlling the size of $\mathcal{F}_t^{(s)}$ will likely involve a quite different approach than the one based on diagonal dominance (in Lemma 10) explored in the present work.

---

[7] The optimum x^* of BP must satisfy $\|x^* - \tilde{x}\|_1 \leq \|x - \tilde{x}\|_1$, so $\|x^* - x\|_1 \leq 2\|x - \tilde{x}\|_1 \leq 2Bs$.

5.3   A Unified and More Natural Approach?

Our solution to Problem 2 has two separate steps: first recovering the subspace that contains all the clean data, and then denoising each single data point with Basis Pursuit using this subspace information. Note that the distributional assumptions (i.e., Gaussianity) on the data are not used in the second step. It is then interesting to ask whether there is a more natural single-phase approach that achieves a similar performance guarantee.

For example, if we already know the low-rank covariance matrix Σ (and thus the subspace U), and the goal is to estimate the mean μ, the natural convex relaxation would be to set variables $\hat{\mu} \in \mathbb{R}^d$ and $y^{(1)}, y^{(2)}, \ldots, y^{(n)} \in U$, and minimize an appropriate combination of the following:

- $\sum_{i=1}^{n} \|\tilde{x}^{(i)} - (\mu + y^{(i)})\|_1$, which is the ℓ_1 surrogate for the sparsity.
- $\sum_{i=1}^{n} [y^{(i)}]^\top \Sigma^{-1} y^{(i)}$, which penalizes the noises $y^{(i)}$ that are unlikely to occur.

Taking one more step back, without directly modeling the Gaussianity of the data, does the "vanilla" robust PCA algorithm succeed for the setups in this paper? The near optimality of BP proved in our work might bring some optimism that the natural convex relaxation approach guarantees not only exact recovery under structural assumptions but also reasonable error bounds under the randomized location model.

References

1 Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, pages 1171–1197, 2012.

2 Khanh Do Ba, Piotr Indyk, Eric Price, and David P Woodruff. Lower bounds for sparse recovery. In *Symposium on Discrete Algorithms (SODA)*, pages 1190–1197, 2010.

3 Ainesh Bakshi and Pravesh K Kothari. List-decodable subspace recovery: Dimension independent error in polynomial time. In *Symposium on Discrete Algorithms (SODA)*, pages 1279–1297, 2021.

4 Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Symposium on Theory of Computing (STOC)*, pages 671–680, 2008.

5 Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

6 Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Transactions on information theory*, 52(2):489–509, 2006.

7 Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Symposium on Theory of Computing (STOC)*, pages 47–60, 2017.

8 Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

9 Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning (ICML)*, pages 774–782, 2013.

10 M Deza and P Frankl. Erdős–ko–rado theorem—22 years later. *SIAM Journal on Algebraic Discrete Methods*, 4(4):419–431, 1983.

11 Ilias Diakonikolas, Daniel Kane, Ankit Pensia, and Thanasis Pittas. Nearly-linear time and streaming algorithms for outlier-robust pca. In *International Conference on Machine Learning (ICML)*, pages 7886–7921, 2023.

12 Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint*, 2019. `arXiv:1911.05911`.

13 Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Symposium on Discrete Algorithms (SODA)*, pages 2745–2754, 2019.

14 David L Donoho. Compressed sensing. *Transactions on information theory*, 52(4):1289–1306, 2006.

15 P. Erdős, Chao Ko, and R. Rado. Intersection theorems for systems of finite sets. *The Quarterly Journal of Mathematics*, 12(1):313–320, 1961.

16 Paul Erdős. A problem on independent r-tuples. *Ann. Univ. Sci. Budapest. Eötvös Sect. Math*, 8:93–95, 1965.

17 Jiashi Feng, Huan Xu, and Shuicheng Yan. Robust pca in high-dimension: a deterministic approach. In *International Conference on Machine Learning (ICML)*, pages 1827–1834, 2012.

18 Simon Foucart and Holger Rauhut. Basis pursuit. *A Mathematical Introduction to Compressive Sensing*, pages 77–110, 2013.

19 Peter Frankl and Andrey Kupavskii. The erdős matching conjecture and concentration inequalities. *Journal of Combinatorial Theory, Series B*, 157:366–400, 2022.

20 P Hall. On representatives of subsets. *Journal of the London Mathematical Society*, 1(1):26–30, 1935.

21 Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory (COLT)*, pages 354–375, 2013.

22 Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *Transactions on Information Theory*, 57(11):7221–7234, 2011.

23 Lunjia Hu and Omer Reingold. Robust mean estimation on highly incomplete data with arbitrary outliers. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1558–1566, 2021.

24 Piotr Indyk and Eric Price. K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance. In *Symposium on Theory of Computing (STOC)*, pages 627–636, 2011.

25 Piotr Indyk, Eric Price, and David P Woodruff. On the power of adaptivity in sparse recovery. In *Foundations of Computer Science (FOCS)*, pages 285–294, 2011.

26 Piotr Indyk and Milan Ruzic. Near-optimal sparse recovery in the l1 norm. In *Foundations of Computer Science (FOCS)*, pages 199–207, 2008.

27 Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15689–15701, 2020.

28 Stasys Jukna. *Extremal combinatorics: with applications in computer science*, volume 571. Springer, 2011.

29 Akshay Kamath and Eric Price. Adaptive sparse recovery with limited adaptivity. In *Symposium on Discrete Algorithms (SODA)*, pages 2729–2744, 2019.

30 Gyula OH Katona. A simple proof of the erdős-chao ko-rado theorem. *Journal of Combinatorial Theory, Series B*, 13(2):183–184, 1972.

31 Weihao Kong, Raghav Somani, Sham Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4683–4696, 2020.

32 Gilad Lerman and Tyler Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018.

33 Zifan Liu, Jong Ho Park, Theodoros Rekatsinas, and Christos Tzamos. On robust mean estimation under coordinate-level corruption. In *International Conference on Machine Learning (ICML)*, pages 6914–6924, 2021.

34 Tyler Maunu and Gilad Lerman. Robust subspace recovery with adversarial outliers. *arXiv preprint*, 2019. `arXiv:1904.03275`.

35   Praneeth Netrapalli, U N Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1107–1115, 2014.

36   Noam Nisan. Pseudorandom bits for constant depth circuits. *Combinatorica*, 11:63–70, 1991.

37   Eric Price and David P Woodruff. (1 + eps)-approximate sparse recovery. In *Foundations of Computer Science (FOCS)*, pages 295–304, 2011.

38   Eric Price and David P Woodruff. Lower bounds for adaptive sparse recovery. In *Symposium on Discrete Algorithms (SODA)*, pages 652–663, 2013.

39   Prasad Raghavendra and Morris Yau. List decodable subspace recovery. In *Conference on Learning Theory (COLT)*, pages 3206–3226, 2020.

40   Vojtěch Rödl. On a packing and covering problem. *European Journal of Combinatorics*, 6(1):69–78, 1985.

41   Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

42   Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust pca: The high-dimensional case. *Transactions on Information Theory*, 59(1):546–572, 2013.

43   Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *Transactions on Information Theory*, 5(58):3047–3064, 2012.

44   Wenzhuo Yang and Huan Xu. A unified framework for outlier-robust pca-like algorithms. In *International Conference on Machine Learning (ICML)*, pages 484–493, 2015.