High Dimensional Structured Superposition Models

Qilong Gu
Dept of Computer Science & Engineering
University of Minnesota, Twin Cities
guxxx396@cs.umn.edu

Arindam Banerjee
Dept of Computer Science & Engineering
University of Minnesota, Twin Cities
banerjee@cs.umn.edu

Abstract

High dimensional superposition models characterize observations using parameters which can be written as a sum of multiple component parameters, each with its own structure, e.g., sum of low rank and sparse matrices, sum of sparse and rotated sparse vectors, etc. In this paper, we consider general superposition models which allow sum of any number of component parameters, and each component structure can be characterized by any norm. We present a simple estimator for such models, give a geometric condition under which the components can be accurately estimated, characterize sample complexity of the estimator, and give high probability non-asymptotic bounds on the componentwise estimation error. We use tools from empirical processes and generic chaining for the statistical analysis, and our results, which substantially generalize prior work on superposition models, are in terms of Gaussian widths of suitable sets.

1 Introduction

For high-dimensional structured estimation problems [3, 15], considerable advances have been made in accurately estimating a sparse or structured parameter $\theta \in \mathbb{R}^p$ even when the sample size n is far smaller than the ambient dimensionality of θ, i.e., $n \ll p$. Instead of a single structure, such as sparsity or low rank, recent years have seen interest in parameter estimation when the parameter θ is a *superposition* or *sum of multiple different structures*, i.e., $\theta = \sum_{i=1}^{k} \theta_i$, where θ_1 may be sparse, θ_2 may be low rank, and so on [1, 6, 7, 9, 11, 12, 13, 23, 24].

In this paper, we substantially generalize the non-asymptotic estimation error analysis for such superposition models such that (i) the parameter θ can be the superposition of *any number of component parameters* θ_i, and (ii) the structure in each θ_i can be captured by *any suitable norm* $R_i(\theta_i)$. We will analyze the following linear measurement based superposition model

$$y = X \sum_{i=1}^{k} \theta_i + \omega \,, \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$ is a random sub-Gaussian design or compressive matrix, k is the number of components, θ_i is one component of the unknown parameters, $y \in \mathbb{R}^n$ is the response vector, and $\omega \in \mathbb{R}^n$ is random noise independent of X. The structure in each component θ_i is captured by any suitable norm $R_i(\cdot)$, such that $R_i(\theta_i)$ has a small value, e.g., sparsity captured by $\|\theta_i\|_1$, low-rank (for matrix θ_i) captured by the nuclear norm $\|\theta_i\|_*$, etc. Popular models such as Morphological Component Analysis (MCA) [10] and Robust PCA [6, 9] can be viewed as a special cases of this framework (see Section D).

The superposition estimation problem can be posed as follows: Given (y, X) generated following (1), estimate component parameters $\{\hat{\theta}_i\}$ such that all the component-wise estimation errors $\Delta_i = \hat{\theta}_i - \theta_i^*$, where θ_i^* is the population mean, are small. Ideally, we want to obtain high-probability non-asymptotic bounds on the total componentwise error measured as $\sum_{i=1}^{k} \|\hat{\theta}_i - \theta_i^*\|_2$, with the bound improving (getting smaller) with increase in the number n of samples.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

We propose the following estimator for the superposition model in (1):

$$\min_{\{\theta_1,\ldots,\theta_k\}} \left\| y - X\sum_{i=1}^{k}\theta_i \right\|_2^2 \qquad \text{s.t.} \quad R_i(\theta_i) \le \alpha_i \ , \quad i = 1,\ldots,k \ , \tag{2}$$

where α_i are suitable constants. In this paper, we focus on the case where $\alpha_i = R_i(\theta_i^*)$, e.g., if θ_i^* is s-sparse with $\|\theta_i^*\|_2 = 1$ and $R_i(\cdot) = \|\cdot\|_1$, then $\alpha_i = \sqrt{s}$ so that $R_i(\theta_i^*) \le \sqrt{s}$, noting that recent advances [16] can be used to extend our results to more general settings.

The superposition estimator in (2) succeeds if a certain geometric condition, which we call *structural coherence* (SC), is satisfied by certain sets (cones) associated with the component norms $R_i(\cdot)$. Since the estimate $\hat{\theta}_i = \theta_i^* + \Delta_i$ is in the feasible set of the optimization problem (2), the error vector Δ_i satisfies the constraint $R_i(\theta_i^* + \Delta_i) \le \alpha_i$ where $\alpha_i = R_i(\theta_i^*)$. The SC condition is a geometric relationship between the corresponding error cones $\mathcal{C}_i = \mathrm{cone}\{\Delta_i | R_i(\theta_i^* + \Delta_i) \le R_i(\theta_i^*)\}$.

If SC is satisfied, then we can show that the sum of componentwise estimation error can be bounded with high probability, and the bound takes the form:

$$\sum_{i=1}^{k} \|\hat{\theta}_i - \theta_i^*\|_2 \le c\frac{\max_i w(\mathcal{C}_i \cap B_p) + \sqrt{\log k}}{\sqrt{n}} \ , \tag{3}$$

where n is the sample size, k is the number of components, and $w(\mathcal{C}_i \cap B_p)$ is the Gaussian width [3, 8, 22] of the intersection of the error cone $\mathcal{C}_i$ with the unit Euclidean ball $B_p \subseteq \mathbb{R}^p$. Interestingly, the estimation error decreases at the rate of $1/\sqrt{n}$, similar to the case of single parameter estimators [15, 3], and depends only logarithmically on the number of components k. Further, while dependency of the error on Gaussian width of the error cone has been established in recent results involving a single parameter [3, 22], the bound in (3) depends on the maximum of the Gaussian width of individual error cones, not their sum. The analysis thus gives a general way to construct estimators for superposition problems along with high-probability non-asymptotic upper bounds on the sum of componentwise errors. To show the generality of our work, we review and compare related work in Appendix B.

Notation: In this paper, we use $\|.\|$ to denote vector norm, and $|||.|||$ to denote operator norm. For example, $\|.\|_2$ is the Euclidean norm for a vector or matrix, and $|||.|||_*$ is the nuclear norm of a matrix. We denote $\mathrm{cone}\{\mathcal{E}\}$ as the smallest closed cone that contains a given set $\mathcal{E}$. We denote $\langle ., . \rangle$ as the inner product.

The rest of this paper is organized as follows: We start with a deterministic estimation error bound in Section 2, while laying down the key geometric and statistical quantities involved in the analysis. In Section 3, we discuss the geometry of the structural coherence (SC) condition, and in Section 4 show that the geometric SC condition implies statistical restricted eigenvalue (RE) condition. In Section 5, we develop the main error bound on the sum of componentwise errors which hold with high probability for sub-Gaussian designs and noise. We apply our error bound to practical problems in Section 6, and present experimental results in Section 7. We conclude in Section 8. In the Appendix, we compare an estimator using "infimal convolution"[18] of norms with our estimator (2) for the noiseless case, and provide some addition examples and experiments. The proofs of all technical results are also in the Appendix.

2 Error Structure and Recovery Guarantees

In this section, we start with some basic results and, under suitable assumptions, provide a deterministic bound for the componentwise estimation error in superposition models. Subsequently, we will show that the assumptions made here hold with high probability as long as a purely geometric non-probabilistic condition characterized by structural coherence (SC) is satisfied.

Let $\{\hat{\theta}_i\}$ be a solution to the superposition estimation problem in (2), $\{\theta_i^*\}$ be the optimal (population) parameters involved in the true data generation process. Let $\Delta_i = \hat{\theta}_i - \theta_i^*$ be the error vector for component i of the superposition. Our goal is to provide a preliminary understanding of the structure of error sets where Δ_i live, identify conditions under which a bound on the total componentwise error $\sum_{i=1}^{k} \|\hat{\theta}_i - \theta_i^*\|_2$ will hold, and provide a preliminary version of such a bound, which will be subsequently refined to the form in (3) in Section 5. Since $\hat{\theta}_i = \theta_i^* + \Delta_i$ lies in the feasible set of (2),

as discussed in Section 1, the error vectors Δ_i will lie in the error sets $\mathcal{E}_i = \{\Delta_i \in \mathbb{R}^p | R_i(\theta_i^* + \Delta_i) \leq R_i(\theta_i^*)\}$ respectively. For the analysis, we will be focusing on the cone of such error sets, given by

$$\mathcal{C}_i = \text{cone}\{\Delta_i \in \mathbb{R}^p | R_i(\theta_i^* + \Delta_i) \leq R_i(\theta_i^*)\} \; . \tag{4}$$

Let $\theta^* = \sum_{i=1}^k \theta_i^*$, $\hat{\theta} = \sum_{i=1}^k \hat{\theta}_i$, and $\Delta = \sum_{i=1}^k \Delta_i$, so that $\Delta = \hat{\theta} - \theta^*$. From the optimality of $\hat{\theta}$ as a solution to (2), we have

$$\|y - X\hat{\theta}\|^2 \leq \|y - X\theta^*\|^2 \; \Rightarrow \; \|X\Delta\|^2 \leq 2\omega^T X\Delta \; , \tag{5}$$

using $\hat{\theta} = \theta^* + \Delta$ and $y = X\theta^* + \omega$. In order to establish recovery guarantees, under suitable assumptions we construct a lower bound to $\|X\Delta\|^2$, the left hand side of (5). The lower bound is a generalized form of the *restricted eigenvalue* (RE) condition studied in the literature [4, 5, 17]. We also construct an upper bound to $\omega^T X\Delta$, the right hand side of (5), which needs to carefully analyze the noise-design (ND) interaction, i.e., between the noise ω and the design X.

We start by assuming that a generalized form of RE condition is satisfied by the superposition of errors: there exists a constant $\kappa > 0$ such that for all $\Delta_i \in \mathcal{C}_i, i = 1, 2, \ldots, k$:

$$\text{(RE)} \qquad \frac{1}{\sqrt{n}} \left\| X \sum_{i=1}^k \Delta_i \right\|_2 \geq \kappa \sum_{i=1}^k \|\Delta_i\|_2 \; . \tag{6}$$

The above RE condition considers the following set:

$$\mathcal{H} = \left\{ \textstyle\sum_{i=1}^k \Delta_i : \Delta_i \in \mathcal{C}_i, \sum_{i=1}^k \|\Delta_i\|_2 = 1 \right\} \; . \tag{7}$$

which involves all the k error cones, and the lower bound is over the sum of norms of the component wise errors. If $k = 1$, the RE condition in (6) above simplifies to the widely studied RE condition in the current literature on Lasso-type and Dantzig-type estimators [4, 17, 3] where only one error cone is involved. If we set all components but Δ_i to zero, then (6) becomes the RE condition only for component i. We also note that the general RE condition as explicitly stated in (6) has been implicitly used in [1] and [24]. For subsequent analysis, we introduce the set $\bar{\mathcal{H}}$ defined as

$$\bar{\mathcal{H}} = \left\{ \textstyle\sum_{i=1}^k \Delta_i : \Delta_i \in \mathcal{C}_i, \sum_{i=1}^k \|\Delta_i\|_2 \leq 1 \right\} . \tag{8}$$

noting that $\mathcal{H} \subset \bar{\mathcal{H}}$.

The general RE condition in (6) depends on the random design matrix X, and is hence an inequality which will hold with certain probability depending on X and the set $\mathcal{H}$. For superposition problems, the probabilistic RE condition as in (6) is intimately related to the following deterministic *structural coherence* (SC) condition on the interaction of the different component cones $\mathcal{C}_i$, without any explicit reference to the random design matrix X: there is a constant $\rho > 0$ such that for all $\Delta_i \in \mathcal{C}_i, i = 1, \ldots, k$,

$$\text{(SC)} \qquad \left\| \sum_{i=1}^k \Delta_i \right\|_2 \geq \rho \sum_{i=1}^k \|\Delta_i\|_2 \; . \tag{9}$$

If $k = 1$, the SC condition is trivially satisfied with $\rho = 1$. Since most existing literature on high-dimensional structured models focus on the $k = 1$ setting [4, 17, 3], there was no reason to study the SC condition carefully. For $k > 1$, the SC condition (9) implies a non-trivial relationship among the component cones. In particular, if the SC condition is true, then the sum $\sum_{i=1}^k \Delta_i$ being zero implies that each component Δ_i must also be zero. As presented in (9), the SC condition comes across as an algebraic condition. In Section 3, we present a geometric characterization of the SC condition [13], and illustrate that the condition is both necessary and sufficient for accurate recovery of each component. In Section 4, we show that for sub-Gaussian design matrices X, the SC condition in (9) in fact implies that the RE condition in (6) will hold with high probability, after the number of samples crosses a certain sample complexity, which depends on the Gaussian width of the component cones. For now, we assume the RE condition in (6) to hold, and proceed with the error bound analysis.

To establish recovery guarantee, following (5), we need an upper bound on the interaction between noise $\boldsymbol{\omega}$ and design X [3, 14]. In particular, we consider the *noise-design* (ND) interaction

$$\text{(ND)} \qquad s_n(\gamma) = \inf_{s>0} \left\{ s : \sup_{u \in s\mathcal{H}} \frac{1}{\sqrt{n}} \omega^T X u \leq \gamma s^2 \sqrt{n} \right\} \; , \tag{10}$$

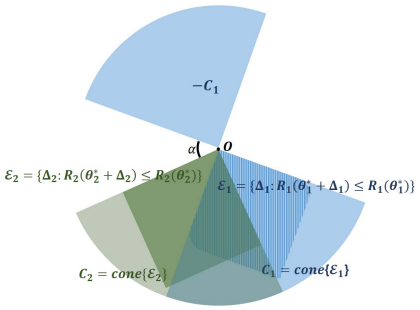


Figure 1: Geometry of SC condition when $k = 2$. The error sets $\mathcal{E}_1$ and $\mathcal{E}_2$ are respectively shown as blue an green squares, and the corresponding error cones are $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively. $-\mathcal{C}_1$ is the reflection of error cone $\mathcal{C}_1$. If $-\mathcal{C}_1$ and $\mathcal{C}_2$ do not share a ray, i.e., the angle α between the cones is larger than 0, then $\delta_0 < 1$, and the SC condition will hold.

where $\gamma > 0$ is a constant, and $s\mathcal{H}$ is the scaled version of $\mathcal{H}$ where the scaling factor is $s > 0$. Here, $s_n(\gamma)$ denotes the minimal scaling needed on $\mathcal{H}$ such that one obtains a uniform bound over $\Delta \in s\mathcal{H}$ of the form: $\frac{1}{n}\omega^T X\Delta \leq \gamma s_n^2(\gamma)$. Then, from the basic inequality in (5), with the bounds implied by the RE condition and the ND interaction, we have

$$\frac{1}{\sqrt{n}}\|X\Delta\|_2 \leq \frac{1}{\sqrt{n}}\sqrt{\omega^T X\Delta} \quad \Rightarrow \quad \kappa \sum_{i=1}^{k} \|\Delta_i\|_2 \leq \sqrt{\gamma} s_n(\gamma) \,, \tag{11}$$

which implies a bound on the component-wise error. The main deterministic bound below states the result formally:

Theorem 1 (Deterministic bound) *Assume that the RE condition in (6) is satisfied in $\mathcal{H}$ with parameter κ. Then, if $\kappa^2 > \gamma$, we have $\sum_{i=1}^{k} \|\Delta_i\|_2 \leq 2s_n(\gamma)$.*

The above bound is deterministic and holds only when the RE condition in (6) is satisfied with constant κ such that $\kappa^2 > \gamma$. In the sequel, we first give a geometric characterization of the SC condition in Section 3, and show that the SC condition implies the RE condition with high probability in Section 4. Further, we give a high probability characterization of $s_n(\gamma)$ based on the noise $\boldsymbol{\omega}$ and design X in terms of the Gaussian widths of the component cones, and also illustrate how one can choose γ in Section 5. With these characterizations, we will obtain the desired component-wise error bound of the form (3).

3 Geometry of Structural Coherence

In this section, we give a geometric characterize the structural coherence (SC) condition in (9). We start with the simplest case of two vectors x, y. If they are not reflections of each other, i.e., $x \neq -y$, then the following relationship holds:

Proposition 2 *If there exists a $\delta < 1$ such that $-\langle x, y\rangle \leq \delta\|x\|_2\|y\|_2$, then*

$$\|x + y\|_2 \geq \sqrt{\tfrac{1-\delta}{2}}(\|x\|_2 + \|y\|_2) \,. \tag{12}$$

Next, we generalize the condition of Proposition 2 to vectors in two different cones $\mathcal{C}_1$ and $\mathcal{C}_2$. Given the cones, define

$$\delta_0 = \sup_{x \in \mathcal{C}_1 \cap \mathcal{S}^{p-1}, y \in \mathcal{C}_2 \cap \mathcal{S}^{p-1}} -\langle x, y\rangle \,. \tag{13}$$

By construction, $-\langle x, y\rangle \leq \delta_0\|x\|_2\|y\|_2$ for all $x \in \mathcal{C}_1$ and $y \in \mathcal{C}_2$. If $\delta_0 < 1$, then (12) continues to hold for all $x \in \mathcal{C}_1$ and $y \in \mathcal{C}_2$ with constant $\sqrt{(1-\delta_0)/2} > 0$. Note that this corresponds to the SC condition with $k = 2$ and $\rho = \sqrt{(1-\delta_0)/2}$. We can interpret this geometrically as follows: first reflect cone $\mathcal{C}_1$ to get $-\mathcal{C}_1$, then δ is the cosine of the minimum angle between $-\mathcal{C}_1$ and $\mathcal{C}_2$. If $\delta_0 = 1$, then $-\mathcal{C}_1$ and $\mathcal{C}_2$ share a ray, and structural coherence does not hold. Otherwise, $\delta_0 < 1$, implying $-\mathcal{C}_1 \cap \mathcal{C}_2 = \{0\}$, i.e., the two cones intersect only at the origin, and structural coherence holds.

For the general case involving k cones, denote

$$\delta_i = \sup_{u \in -\mathcal{C}_i \cap S_{p-1}, v \in \sum_{j \neq i} \mathcal{C}_j \cap S_{p-1}} \langle u, v\rangle \,. \tag{14}$$

In recent work, [13] concluded that if $\delta_i < 1$ for each $i = 1, \ldots, k$ then $-\mathcal{C}_i$ and $\sum_{j \neq i} \mathcal{C}_j$ does not share a ray, and the original signal can be recovered in noiseless case. We show that the condition above in fact implies $\rho > 0$ for the SC condition in (9), which is sufficient for accurate recovery even in the noisy case. In particular, with $\delta := \max_i \delta_i$, we have the following result:

Theorem 3 (Structural Coherence (SC) Condition) *Let $\delta := \max_i \delta_i$ with δ_i as defined in (14). If $\delta < 1$, then there exists a $\rho > 0$ such that for any $\Delta_i \in \mathcal{C}_i, i = 1, \ldots, k$, the SC condition in (9) holds, i.e.,*

$$\left\| \sum_{i=1}^k \Delta_i \right\|_2 \geq \rho \sum_{i=1}^k \|\Delta_i\|_2 \, . \tag{15}$$

Thus, the SC condition is satisfied in the general case as long as the reflection $-\mathcal{C}_i$ of any cone $\mathcal{C}_i$ does not intersect, i.e., share a ray, with the Minkowski sum $\sum_{j \neq i} \mathcal{C}_j$ of the other cones.

4 Restricted Eigenvalue Condition for Superposition Models

Assuming that the SC condition is satisfied by the error cones $\{\mathcal{C}_i\}, i = 1, \ldots, k$, in this section we show that the general RE condition in (6) will be satisfied with high probability when the number of samples n in the sub-Gaussian design matrix $X \in \mathbb{R}^{n \times p}$ crosses the sample complexity n_0. We give a precise characterization of the sample complexity n_0 in terms of the Gaussian width of the set $\mathcal{H}$.

Our analysis is based on the results and techniques in [20, 14], and we note that [3] has related results using mildly different techniques. We start with a restricted eigenvalue condition on $\mathcal{C}$. For a random vector $Z \in \mathbb{R}^p$, we define marginal tail function for an arbitrary set E as

$$Q_\xi(E; Z) = \inf_{u \in E} P(|\langle Z, u \rangle| \geq \xi) \, , \tag{16}$$

noting that it is deterministic given the set $E \subseteq \mathbb{R}^p$. Let $\epsilon_i, i = 1, \ldots, n$, be independent Rademacher random variables, i.e., random variable with probability $\frac{1}{2}$ of being either $+1$ or -1, and let $X_i, i = 1, \ldots, n$, be independent copies of Z. We define empirical width of E as

$$W_n(E; Z) = \sup_{u \in E} \langle h, u \rangle, \quad \text{where} \quad h = \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n \epsilon_i X_i \, . \tag{17}$$

With this notation, we recall the following result from [20]:

Lemma 1 *Let $X \in \mathbb{R}^{n \times p}$ be a random design matrix with each row the independent copy of sub-Gaussian random vector Z. Then for any $\xi, \rho, t > 0$, we have*

$$\inf_{u \in \mathcal{H}} \|Xu\|_2 \geq \rho \xi \sqrt{n} Q_{2\rho\xi}(\mathcal{H}; Z) - 2 W_n(\mathcal{H}; Z) - \rho \xi t \tag{18}$$

with probability at least $1 - e^{-\frac{t^2}{2}}$.

In order to obtain lower bound of κ in RE condition (6), we need to lower bound $Q_{2\rho\xi}(\mathcal{H}; Z)$ and upper bound $W_n(\mathcal{H}; Z)$. To lower bound $Q_{2\rho\xi}(\mathcal{H}; Z)$, we consider the spherical cap

$$\mathcal{A} = (\textstyle\sum_{i=1}^k \mathcal{C}_i) \cap \mathcal{S}^{p-1} \, . \tag{19}$$

From [20, 14], one can obtain a lower bound to $Q_\xi(\mathcal{A}; Z)$ based on the Paley-Zygmund inequality. The Paley-Zygmund inequality lower bound the tail distribution of a random variable by its second momentum. Let u be an arbitrary vector, we use the following version of the inequality.

$$P(|\langle Z, u \rangle| \geq 2\xi) \geq \tfrac{[\mathbb{E}|\langle Z, u \rangle| - 2\xi]_+^2}{\mathbb{E}|\langle Z, u \rangle|^2} \tag{20}$$

In the current context, the following result is a direct consequence of SC condition, which shows that $Q_{2\rho\xi}(\mathcal{H}; Z)$ is lower bounded by $Q_\xi(\mathcal{A}; Z)$, which in turn is strictly bounded away from 0 . The proof of Lemma 2 is given in Appendix H.1.

Lemma 2 *Let sets $\mathcal{H}$ and $\mathcal{A}$ be as defined in (7) and (19) respectively. If the SC condition in (9) holds, then the marginal tail functions of the two sets have the following relationship:*

$$Q_{\rho\xi}(\mathcal{H}; Z) \geq Q_\xi(\mathcal{A}; Z). \tag{21}$$

Next we discuss how to upper bound the empirical width $W_n(\mathcal{H}; Z)$. Let set $\mathcal{E}$ be arbitrary, and random vector $g \sim \mathcal{N}(0, I_p)$ be a standard Gaussian random vector in $\mathbb{R}^p$. The Gaussian width [3] of $\mathcal{E}$ is defined as

$$w(E) = \mathbb{E} \sup_{u \in \mathcal{E}} \langle g, u \rangle. \tag{22}$$

Empirical width $W_n(\mathcal{H}; Z)$ can be seen as the supremum of a stochastic process. One way to upper bound the supremum of a stochastic process is by generic chaining [19, 3, 20], and by using generic

chaining we can upper bound the stochastic process by a Gaussian process, which is the Gaussian width.

As we can bound $Q_{2\rho\xi}(\mathcal{H}; Z)$ and $W_n(\mathcal{H}; Z)$, we come to the conclusion on RE condition. Let $X \in \mathbb{R}^{n\times p}$ be a random matrix where each row is an independent copy of the sub-Gaussian random vector $Z \in \mathbb{R}^p$, and where Z has sub-Gaussian norm $\|Z\|_{\psi_2} \leq \sigma_x$ [21]. Let $\alpha = \inf_{u\in\mathcal{S}^{p-1}} \mathbb{E}[|\langle Z, u\rangle|]$ so that $\alpha > 0$ [14, 20]. We have the following lower bound of the RE condition. The proof of Theorem 4 is based on the proof of [20, Theorem 6.3], and we give it in appendix H.2.

Theorem 4 (Restricted Eigenvalue Condition) *Let X be the sub-Gaussian design matrix that satisfies the assumptions above. If the SC condition (9) holds with a $\rho > 0$, then with probability at least $1 - \exp(-t^2/2)$, we have*

$$\inf_{u\in\mathcal{H}} \|Xu\|_2 \geq c_1\rho\sqrt{n} - c_2 w(\mathcal{H}) - c_3\rho t \tag{23}$$

where c_1, c_2 and c_3 are positive constants determined by σ_x, σ_ω and α.

To get a $\kappa > 0$ in (6), one can simply choose $t = (c_1\rho\sqrt{n} - c_2 w(\mathcal{H}))/2c_3\rho$. Then as long as $n > c_4 w^2(\mathcal{H})/\rho^2$ for $c_4 = c_2^2/c_1^2$, we have

$$\kappa = \inf_{u\in\mathcal{H}} \tfrac{1}{\sqrt{n}}\|Xu\|_2 \geq \tfrac{1}{2}\left(c_1\rho - c_2\tfrac{w(\mathcal{H})}{\sqrt{n}}\right) > 0,$$

with high probability.

From the discussion above, if SC condition holds and the sample size n is large enough, then we can find a matrix X such that RE condition holds. On the other hand, once there is a matrix X such that RE condition holds, then we can show that SC must also be true. Its proof is give in Appendix H.3.

Proposition 5 *If X is a matrix such that the RE condition* (6) *holds for $\Delta_i \in \mathcal{C}_i$, then the SC condition (9) holds.*

Proposition 5 demonstrates that SC condition is a necessary condition for the possibility of RE. If SC condition does not hold, then there is $\{\Delta_i\}$ such that $\Delta_i \neq 0$ for some $i = 1, \ldots, k$, but $\|\sum_{i=1}^k \Delta_i\|_2 = 0$ which implies $\sum_{i=1}^k \Delta_i = 0$. Then for every matrix X, we have $X\sum_{i=1}^k \Delta_i = 0$, and RE condition is not possible.

5 General Error Bound

Recall that the error bound in Theorem 1 is given in terms of the noise-design (ND) interaction

$$s_n(\gamma) = \inf_{s>0}\left\{s : \sup_{u\in s\mathcal{C}} \tfrac{1}{\sqrt{n}}\omega^T Xu \leq \gamma s^2\sqrt{n}\right\}. \tag{24}$$

In this section, we give a characterization of the ND interaction, which yields the final bound on the componentwise error as long as $n \geq n_0$, i.e., the sample complexity is satisfied.

Let ω be a centered sub-Gaussian random vector, and its sub-Gaussian norm $\|\omega\|_{\psi_2} \leq \sigma_\omega$. Let X be a row-wise i.i.d. sub-Gaussian random matrix, for each row Z, its sub-Gaussian norm $\|Z\|_{\psi_2} \leq \sigma_x$. The ND interaction can be bounded by the following conclusion, and the proof of lemma 3 is given in appendix I.1.

Lemma 3 *Let design $X \in \mathbb{R}^{n\times p}$ be a row-wise i.i.d. sub-Gaussian random matrix, and noise $\omega \in \mathbb{R}^n$ be a centered sub-Gaussian random vector. Then $s_n(\gamma) \leq c\frac{w(\bar{\mathcal{H}})}{\gamma\sqrt{n}}$. for some constant $c > 0$ with probability at least $1 - c_1\exp(-c_2 w^2(\bar{\mathcal{H}})) - c_3\exp(-c_4 n)$. Constant c depends on σ_x and σ_ω.*

In lemma 3 and theorem 6, we need the Gaussian width of $\bar{\mathcal{H}}$ and $\mathcal{H}$ respectively. From definition, both $\bar{\mathcal{H}}$ and $\mathcal{H}$ is related to the union of different cones; therefore bounding the width of $\bar{\mathcal{H}}$ and $\mathcal{H}$ may be difficult. We have the following bound of $w(\mathcal{H})$ and $w(\bar{\mathcal{H}})$ in terms of the width of the component spherical caps. The proof of Lemma 4 is given in Appendix I.2.

Lemma 4 (Gaussian width bound) *Let $\mathcal{H}$ and $\bar{\mathcal{H}}$ be as defined in (7) and (8) respectively. Then, we have $w(\mathcal{H}) = O\left(\max_i w(\mathcal{C}_i \cap \mathcal{S}_{p-1}) + \sqrt{\log k}\right)$ and $w(\bar{\mathcal{H}}) = O\left(\max_i w(\mathcal{C}_i \cap B_p) + \sqrt{\log k}\right)$.*

By applying lemma 4, we can derive the error bound using the Gaussian width of individual error cone. From our conclusion on deterministic bound in theorem 1, we can choose an appropriate γ such that $\kappa^2 > \gamma$. Then, by combining the result of theorem 1, theorem 4, lemma 3 and lemma 4, we have the final form of the bound, as originally discussed in (3):

Theorem 6 *For estimator (2), let $\mathcal{C}_i = \text{cone}\{\Delta : R_i(\theta_i^* + \Delta) \leq R_i(\theta_i^*)\}$, design X be a random matrix with each row an independent copy of sub-Gaussian random vector Z, noise ω be a centered sub-Gaussian random vector, and $B_p \subseteq \mathbb{R}^p$ be the centered unit euclidean ball. If sample size $n > c(\max_i w^2(\mathcal{C}_i \cap S_{p-1}) + \log k)/\rho^2$, then we have with probability at least $1 - \frac{\eta_1}{k}\exp(-\eta_2 \max_i w^2(\mathcal{C}_i \cap S_{p-1})) - \eta_3 \exp(-\eta_4 n)$,*

$$\textstyle\sum_{i=1}^k \|\hat{\theta}_i - \theta_i^*\|_2 \leq C\frac{\max_i w(\mathcal{C}_i \cap B_p) + \sqrt{\log k}}{\rho^2\sqrt{n}}, \tag{25}$$

for constants $c, C > 0$ that depend on sub-Gaussian norms $\|\|Z\|\|_{\phi_2}$ and $\|\|\omega\|\|_{\phi_2}$.

Thus, assuming the SC condition in (9) is satisfied, the sample complexity and error bound of the estimator depends on the largest Gaussian width, rather than the sum of Gaussian widths. The result can be viewed as a direct generalization of existing results for $k = 1$, when the SC condition is always satisfied, and the sample complexity and error is given by $w^2(\mathcal{C}_1 \cap S_{p-1})$ and $w(\mathcal{C}_1 \cap B_p)$ [3, 8].

6 Application of General Bound

In this section, we instantiate the general error bounds on Morphological Component Analysis (MCA), and low-rank and sparse matrix decomposition. The comprehensive results are provided in Appendix D.

6.1 Morphological Component Analysis

In Morphological Component Analysis [10], we consider the following linear model

$$y = X(\theta_1^* + \theta_2^*) + \omega \tag{26}$$

where vector θ_1^* is sparse and θ_2^* is sparse under a rotation Q. Consider the following estimator

$$\min_{\theta_1,\theta_2} \|y - X(\theta_1 + \theta_2)\|_2^2 \quad s.t. \quad \|\theta_1\|_1 \leq \|\theta_1^*\|_1, \|Q\theta_2\|_1 \leq \|Q\theta_2^*\|_1, \tag{27}$$

where vector $y \in \mathbb{R}^n$ is the observation, vectors $\theta_1, \theta_2 \in \mathbb{R}^p$ are the parameters we want to estimate, matrix $X \in \mathbb{R}^{n\times p}$ is a sub-Gaussian random design, matrix $Q \in \mathbb{R}^{p\times p}$ is orthogonal. We assume θ_1^* and $Q\theta_2^*$ are s_1-sparse and s_2-sparse vectors respectively. Function $\|Q.\|_1$ is still a norm. In general, we can derive the following error bound from Theorem 6:

$$\|\theta_1 - \theta_1^*\|_2 + \|\theta_2 - \theta_2^*\|_2 = O\left(\max\left\{\sqrt{\tfrac{s_1 \log p}{n}}, \sqrt{\tfrac{s_2 \log p}{n}}\right\}\right).$$

6.2 Low-rank and Sparse Matrix Decomposition

To recover a sparse matrix and low-rank matrix from their sum [6, 9], one can use L1 norm to induce sparsity and nuclear norm to induce low-rank. These two kinds of norm ensure that the sparsity and the rank of the estimated matrices are small. Suppose we have a rank-r matrix L^* and a sparse matrix S^* with s nonzero entries, $S^*, L^* \in R^{d_1\times d_2}$. Our observation Y comes from the following problem

$$Y_i = \langle X_i, L^* + S^*\rangle + E_i, i = 1, \ldots, n,$$

where each $X_i \in \mathbb{R}^{d_1\times d_2}$ is a sub-Gaussian random design matrix. E_i is the noise matrix. The estimator takes the form:

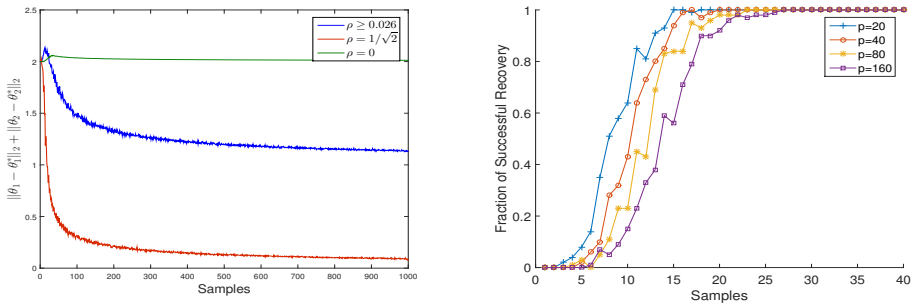
$$\min_{L,S} \sum_{i=1}^n (Y_i - \langle X_i, L + S\rangle)^2 \quad s.t. \quad \|\|L\|\|_* \leq \|\|L^*\|\|_*, \|S\|_1 \leq \|S^*\|_1. \tag{28}$$

By using Theorem 6, and existing results on Gaussian widths, the error bound is given by

$$\|L - L^*\|_2 + \|S - S^*\|_2 = O\left(\max\left\{\sqrt{\tfrac{s\log(d_1 d_2)}{n}}, \sqrt{\tfrac{r(d_1+d_2-r)}{n}}\right\}\right).$$

7 Experimental Results

In this section, we confirm the theoretical results in this paper with some simple experiments. We show our experimental results under different settings. In our experiments we focus on MCA when $k = 2$. The design matrix X are generated from Gaussian distribution such that every entry of X



(a) (b)

Figure 2: (a) Effect of parameter ρ on estimation error when noise $\omega \neq 0$. We choose the parameter ρ to be 0, $1/\sqrt{2}$, and a random sample. (b) Effect of dimension p on fraction of successful recovery in noiseless case. Dimension p varies in $\{20, 40, 50, 150\}$

subjects to $\mathcal{N}(0, 1)$. The noise ω is generated from Gaussian distribution such that every entry of ω subjects to $\mathcal{N}(0, 1)$. We implement our algorithm 1 in MATLAB. We use synthetic data in all our experiments, and let the true signal

$$\theta_1 = (\underbrace{1, \ldots, 1}_{s_1}, 0 \ldots, 0), Q\theta_2 = (\underbrace{1, \ldots, 1}_{s_2}, 0 \ldots, 0)$$

We generate our data in different ways for our three experiments.

7.1 Recovery From Noisy Observation

In our first experiment, we test the impact of ρ on the estimation error. We choose three different matrices Q, and ρ is determined the choice of Q. The first Q is given by random sampling: we sample a random orthogonal matrix Q such that $Q_{ij} > 0$, and ρ is lower bounded by (42). The second and third Q is given by identity matrix I and its negative $-I$; therefore $\rho = 1/\sqrt{2}$ and $\rho = 0$ respectively. We choose dimension $p = 1000$, and let $s_1 = s_2 = 1$. The number of samples n varied between 1 and 1000. Observation y is given by $y = X(\theta_1^* + \theta_2^*) + \omega$. In this experiment, given Q, for each n, we generate 100 pairs of X and w. For each (X, w) pair, we get a solution $\hat{\theta}_1$ and $\hat{\theta}_2$. We take the average over all $\|\hat{\theta}_1 - \theta_1^*\|_2 + \|\hat{\theta}_2 - \theta_2^*\|_2$. Figure 2(a) shows the plot of number of samples vs the average error. From figure 2(a), we can see that the error curve given by random Q lies between curves given by two extreme cases, and larger ρ gives lower curve. In Appendix E, we provide an additional experiment using k-support norm [2].

7.2 Recovery From Noiseless Observation

In our second experiment, we test how the dimension p affects the successful recovery of true value. In this experiment, we choose different dimension p with $p = 20$, $p = 40$, $p = 80$, and $p = 160$. We let $s_1 = s_2 = 1$. To avoid the impact of ρ, for each sample size n, we sample 100 random orthogonal matrices Q. Observation y is given by $y = X(\theta_1^* + \theta_2^*)$. For each solution $\hat{\theta}_1$ and $\hat{\theta}_2$ of (41), we calculate the proportion of Q such that $\|\hat{\theta}_1 - \theta_1^*\|_2 + \|\hat{\theta}_2 - \theta_2^*\|_2 \leq 10^{-4}$. We increase n from 1 to 40, and the plot we get is figure 2(b). From figure 2(b) we can find that the sample complexity required to recover θ_1^* and θ_2^* increases with dimension p.

8 Conclusions

We present a simple estimator for general superposition models and give a purely geometric characterization, based on structural coherence, of when accurate estimation of each component is possible. Further, we establish sample complexity of the estimator and upper bounds on componentwise estimation error and show that both, interestingly, depend on the largest Gaussian width among the spherical caps induced by the error cones corresponding to the component norms. Going forward, it will be interesting to investigate specific component structures which satisfy structural coherence, and also extend our results to allow more general measurement models.

Acknowledgements: The research was also supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS- 1314560, IIS-0953274, IIS-1029711, NASA grant NNX12AQ39A, and gifts from Adobe, IBM, and Yahoo.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

[2] A. Argyriou, R. Foygel, and N. Srebro. Sparse Prediction with the k-Support Norm. In *Advances in Neural Information Processing Systems*, Apr. 2012.

[3] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with Norm Regularization. In *Advances in Neural Information Processing Systems*, 2014.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[5] P. Buhlmann and S. van de Geer. *Statistics for High Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):1–37, 2011.

[7] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

[8] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12:805–849, 2012.

[9] V. Chandrasekaran, S. Sanghavi, P. a. Parrilo, and A. S. Willsky. Rank-Sparsity Incoherence for Matrix Decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[10] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[11] R. Foygel and L. Mackey. Corrupted Sensing: Novel Guarantees for Separating Structured Signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, Feb. 2014.

[12] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

[13] M. B. McCoy and J. A. Tropp. The achievable performance of convex demixing. *arXiv*, 2013.

[14] S. Mendelson. Learning without concentration. *J. ACM*, 62(3):21:1–21:25, June 2015.

[15] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, Nov. 2012.

[16] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp Time–Data Tradeoffs for Linear Inverse Problems. *ArXiv e-prints*, July 2015.

[17] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted Eigenvalue Properties for Correlated Gaussian Designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[18] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[19] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. A Series of Modern Surveys in Mathematics. Springer-Verlag Berlin Heidelberg, 2014.

[20] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. *arXiv*, May 2014.

[21] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, Cambridge, Nov. 2012.

[22] R. Vershynin. Estimation in high dimensions: a geometric perspective. *Sampling Theory, a Renaissance*, pages 3–66, 2015.

[23] J. Wright, A. Ganesh, K. Min, and Y. Ma. Compressive principal component pursuit. *IEEE International Symposium on Information Theory*, pages 1276–1280, 2012.

[24] E. Yang and P. Ravikumar. Dirty statistical models. *Advances in Neural Information Processing Systems*, pages 1–9, 2012.