Appendix

A Accelerated Proximal Algorithm

Algorithm 1 Accelerated Proximal Algorithm

Inputs: X, $\mathbf{y}$.
Initialize: $\{\theta_i^0\}_{i=1}^k = \{0\}$, $\alpha_0 = 0$, $\eta > 0$, $0 < \beta < 1$.
for $t = 0, 1, \ldots, T$ **do**
  Set $\eta^{t+1} = \eta$.
  while true **do**
    for $i = 1, \ldots, k$ **do**
      $\tilde{\theta}_i^{t+1} = \Pi_{\Omega_i}(\theta_i^t - \eta^{t+1}\nabla f_{\theta_i}(\theta^t))$
    end for
    if $f(\tilde{\theta}^{t+1}) \leq f(\theta^t) + \nabla^T f(\theta^t)(\tilde{\theta}^{t+1} - \theta^t) + \frac{1}{2\eta^{t+1}}(\sum_{i=1}^k \|\tilde{\theta}_i^{t+1} - \theta_i^t\|_2^2)$ **then**
      break
    end if
    $\eta^{t+1} = \beta\eta^{t+1}$
  end while
  $\alpha_{t+1} = \frac{1+\sqrt{1+4\alpha_t^2}}{2}$, $\theta_i^{t+1} = \tilde{\theta}_i^{t+1} + \frac{\alpha_t - 1}{\alpha_{t+1}}(\tilde{\theta}_i^{t+1} - \theta_i^t)$
end for

In this section, we propose a general purpose algorithm for solving problem (2). For convenience, with $\theta = \sum_{i=1}^k \theta_i$, we set $f(\theta) = f(\sum_{i=1}^k \theta_i) = \|\mathbf{y} - X\theta\|_2^2$ and $\Omega_i = \{\theta_i | R_i(\theta_i) \leq R_i(\theta_i^*)\}$. While the norms $R_i(.)$ may be non-smooth, one can design a general algorithm as long as the proximal operators $\Pi_{\Omega_i}(v) = \text{argmin}_{u \in \Omega_i} \|u - v\|_2^2$ for each set Ω_i can be efficiently computed. The algorithm is simply the proximal gradient method [23], where each component θ_i is cyclically updated in each iteration (see Algorithm 1):

$$\begin{aligned} \tilde{\theta}_i^{t+1} &= \underset{\theta_i \in \Omega_i}{\text{argmin}} \ \langle \nabla_{\theta_i} f(\theta^t), \theta_i - \theta_i^t \rangle + \frac{1}{2\eta^{t+1}} \|\theta_i - \theta_i^t\|_2^2 \,, \\ &= \underset{\theta_i \in \Omega_i}{\text{argmin}} \ \|\theta_i - (\theta_i^t - \eta^{t+1}\nabla_{\theta_i} f(\theta^t))\|_2^2 \,, \end{aligned} \tag{29}$$

where η^{t+1} is the learning rate. To determine a proper η^{t+1}, we use a backtracking step [4]. Starting from a constant $\eta^{t+1} = \eta$, in each step we first update $\tilde{\theta}_i^{t+1}$; then we decide whether $\tilde{\theta}_i^{t+1}$ satisfies condition:

$$f(\tilde{\theta}^{t+1}) \leq f(\theta^t) + \nabla^T f(\theta^t)(\tilde{\theta}^{t+1} - \theta^t) + \frac{1}{2\eta^{t+1}}(\sum_{i=1}^k \|\tilde{\theta}_i^{t+1} - \theta_i^t\|_2^2). \tag{30}$$

If the condition (30) does not hold, then we decrease η^{t+1} till (30) is satisfied. Based on existing results [4], the basic method can be accelerated by setting the starting point of the next iteration θ_i^{t+1} as a proper combination of $\tilde{\theta}_i^{t+1}$ and θ_i^t. By [4], one can use the updates:

$$\theta_i^{t+1} = \tilde{\theta}_i^{t+1} + \frac{\alpha_t - 1}{\alpha_{t+1}}(\tilde{\theta}_i^{t+1} - \theta_i^t) \,, \quad \text{where} \quad \alpha_{t+1} = \frac{1 + \sqrt{1 + 4\alpha_t^2}}{2} \,. \tag{31}$$

Convergence of Algorithm 1 has been studied in [4]. The backtracking step ensures that the convergence of algorithm 1. The work [4] also give the convergence rate of Algorithm 1, which is $O(1/t^2)$. Therefore, we can always reach a stationary point of problem (2) using Algorithm 1.

B Related Work

Structured superposition models have been studied in recent literatures. Early work focus on the case when k=2 and noise $\omega = 0$, and assume specific structures such as sparse+sparse [14], and low-rank+sparse [11]. [16] analyze error bound for low-rank and sparse matrix decomposition with noise. Recent work have considered more generalized models and structures. [1] analyze the

decomposition of a low-rank matrix plus another matrix with generalized structure. [15] propose an estimator for the decomposition of two generalize structured matrices, while one of them has a random rotation. Because of the increase in practical application and non-trivial of such problem, people have begun to work on unified frameworks for superposition model. In [30], the authors generalize the noiseless matrix decomposition problem to arbitrary number of superposition under random orthogonal measurement. [31] consider the superposition of structures of structures captured by decomposable norm, while [19] consider general norms but with a different measurement model, involving componentwise random rotations. These two papers are similar in spirit to our work, so we briefly discuss and differentiate our work from these papers.

[31] consider a general framework for superposition model, and give a high-probability bound for the following estimation problem:

$$\min_{\theta_i, i=1,\ldots,k} \left\| y - X \sum_{i=1}^{k} \theta_i \right\|_2^2 + \sum_{i=1}^{k} \lambda_i R_i(\theta_i) \tag{32}$$

they assume each $R_i()$ to be a special kind of norm called decomposable norm. the authors used a different approach for RE condition. They decompose $\|X \sum_{i=1}^{k} \Delta_i\|_2$ into two parts. One is

$$\tfrac{1}{n}\|X\Delta_i\|_2^2 \geq \kappa \|\Delta_i\|_2^2, \tag{33}$$

which characterizes the restricted eigenvalue of each error cone. The other is

$$\tfrac{2}{n}|\sum_{i<j}\langle X\Delta_i, X\Delta_j\rangle| \leq \tfrac{\kappa}{2}\sum_{i=1}^{k}\|\Delta_i\|_2^2, \tag{34}$$

which characterizes the interaction between different error cones. (34) is a strong assumption, and RE condition can hold without it. If Δ_i and Δ_j are positively correlated, then large interaction terms will make our RE condition stronger. Therefore their results are restricted.

[18] consider an estimator like (2), which is

$$\min_{\theta_i, i=1,\ldots,k} \left\| y - X \sum_{i=1}^{k} Q_i \theta_i \right\|_2^2 \text{ s.t. } R_i(\theta_i) \leq R_i(\theta_i^*),\ i = 1, \ldots, k, \tag{35}$$

where Q_i are known random rotations. Problem (35) is then transformed into a geometric problem: whether k random cones intersect. The componentwise random rotation can ensure that any kind of combination can be recovered with high probability. However, in practical problems, we need not have such random rotations available as part of the measurements. Further, their analysis is primarily focused on the noiseless case.

C Noiseless Case: Comparing Estimators

In this section, we present a comparative analysis of estimator

$$\min_{\{\theta_i\}} \textstyle\sum_{i=1}^{k} \lambda_i R_i(\theta_i) \quad s.t. \quad X \sum_{i=1}^{k} \theta_i = y \tag{36}$$

with the proposed estimator (2) in the noiseless case, i.e., $\omega = 0$. In essence, we show that the two estimators have similar recovery conditions, but the existing estimator (36) needs additional structure for unique decomposition of θ into the components $\{\hat{\theta}_i\}$.

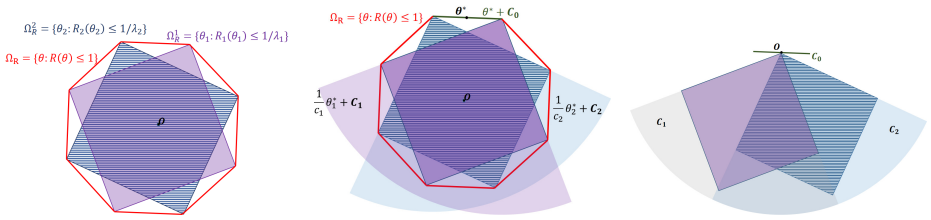
The estimator (36) needs to consider the so-called "infimal convolution" [25, 31] over different norms to get a (unique) decomposition of θ in terms of the components $\{\hat{\theta}_i\}$. Denote

$$R(\theta) = \min_{\{\theta_i\}:\sum_i \theta_i = \theta} \textstyle\sum_{i=1}^{k} \lambda_i R_i(\theta_i)\ . \tag{37}$$

Results in [25] show that (37) is also a norm. Thus estimator (36) can be rewritten as

$$\min_{\theta} R(\theta)\ s.t.\ X\theta = y. \tag{38}$$

Interestingly, the above discussion separates the estimation problem in (36) into two parts—solving (38) to get $\hat{\theta}$, and then solving (37) to get the components $\{\hat{\theta}_i\}$. The problem (38) is a simple structured recovery problem, and is well studied [10, 27]. Using infimal convolution based decomposition problem (37) to get the components $\{\hat{\theta}_i\}$ will be our focus in the sequel.



(a) (b) (c)

Figure 3: **(a)** The relationship of different norm balls when $k = 2$. The blue and purple polygons are the norm ball of norms $R_1(.)$ and $R_2(.)$ respectively. The red line is the outline of $R(.)$ norm ball. Note that for any point in the red line, we will be able to decompose it to the two vertexes around it. Consider the case when $k = 2$. Let $c_i = \lambda_i R_i(\theta_i)$ for $i = 1, 2$. **(b)** is the structure of error around the true value θ^*. The green segment $\mathcal{C}_0$ is a subspace determined by θ_1^* and θ_2^*. For the superposition in (a), error of θ^* is composed of three parts: $\theta^* + \mathcal{C}_0$, $\frac{1}{c_1}\theta_1^* + \mathcal{C}_1$ and $\frac{1}{c_2}\theta_2^* + \mathcal{C}_2$. In **(c)**, we move the green segment and two error cones to the origin, then the uniquely recovery condition is that if we reflect one of the three structures, their intersection remains $\{0\}$.

To get some properties of decomposition (37), we consider the unit norm balls for norm $R(.)$ and component norms $R_i(.)$:

$$\Omega_R = \{\theta \in \mathbb{R}^p : R(\theta) \le 1\} \quad \text{and} \quad \Omega_R^i = \{\theta_i \in \mathbb{R}^p : R_i(\theta_i) \le 1\}, i = 1, \ldots, k .$$

The norm balls are related by the following result, we give the proof in appendix K.1.

Lemma 5 *For a given set $\{\lambda_i\}$, the infimal convolution norm ball Ω_R is the convex hull of $\bigcup_{i=1}^k \frac{1}{\lambda_i}\Omega_R^i$, i.e., $\Omega_R = \mathrm{conv}(\bigcup_{i=1}^k \frac{1}{\lambda_i}\Omega_R^i)$.*

Lemma 5 illustrates what the decomposition (37) should be like. If θ is a point on the surface of the norm ball Ω_R, then the value of $R(\theta)$ is the convex combination of some θ_i on the surface of $\frac{1}{\lambda_i}\Omega_R^i$ such that $R_i(\theta_i) = R(\theta)$. Hence if θ can be successfully decomposed into different components along the direction of θ_i, then we should be able to connect θ_i and θ by a surface on the $R(.)$ norm ball, or they have to be "close". Interestingly, the above intuition of "closeness" between different components θ_i can be described in the language of cones, in a way similar to the structural coherence property discussed in Section 3.

Given the intuition above, we state the main result below. Proof is given in Appendix K.2.

Theorem 7 *Given $\hat{\theta}_1, \ldots, \hat{\theta}_k$ and define*

$$\mathcal{C}_0 = \left\{ \sum_{\theta_i \neq 0} \left(\frac{c_i'}{c_i} - 1 \right) \theta_i \mid c_i' \ge 0, \sum_{i=1}^k c_i' = 1 \right\}. \tag{39}$$

Suppose $dim(span\{\theta_i\}) = k$, then there exist $\lambda_1, \ldots, \lambda_k$ such that $\sum_{i=1}^k \hat{\theta}_i = \theta$ are unique solutions of (37) if and only if there are $c_1, \ldots, c_k$ with $c_i \ge 0$ and $\sum_{i=1}^k c_i = 1$ such that for the corresponding error cone $\mathcal{C}_i$ of $\hat{\theta}_i$ and C_0 defined above, $-\mathcal{C}_i \cap \sum_{j \neq i} \mathcal{C}_j = \{0\}$, for $i = 0, 1, \ldots, k$.

Theorem 7 illustrate that the successful decomposition of (37) requires an additional condition, i.e., $-\mathcal{C}_0 \cap \sum_{i=1}^k \mathcal{C}_i = \{0\}$ beyond that is needed by the SC condition (see Section 3). The additional condition needs us to choose parameters $\{\lambda_i\}$ properly. Theorem 7 shows that $\{\lambda_i\}$ depends on both $\{\theta_i^*\}$ and $\{c_i\}$. For appropriate $\{\theta_i^*\}$, there may be a range of $\{c_i\}$ such that the solution is unique. Therefore, in noiseless situation, if we know $\{R_i(\theta_i^*)\}$, then solving estimator (36) would be a better idea, because it requires less condition to recover the true value and we do not need to choose parameters $\{\lambda_i\}$.

D Examples

In this section, we instantiate the general error bounds on concrete problems, the proofs are provided in appendix J.

D.1 Morphological Component Analysis Using L1 Norm

In Morphological Component Analysis [14], we want to separate a sparse vector θ_1 and another vector θ_2 which is sparse under a rotation Q from their sum. In [14], the authors introduced a quantity

$$M = \max_{i,j} |Q_{ij}|. \tag{40}$$

For small enough M, if the sum of their sparsity is lower than a constant related to M, we can recovery them. We show that for two given sparse vectors, our SC condition is more general.

Consider the following estimator

$$\min_{\theta_1,\theta_2} \|y - X(\theta_1+\theta_2)\|_2^2 \quad s.t. \quad \|\theta_1\|_1 \le \|\theta_1^*\|_1, \|Q\theta_2\|_1 \le \|Q\theta_2^*\|_1, \tag{41}$$

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. It is easy to see that $\|Q.\|_1$ is still a norm. Suppose $s_1 = 1$, $s_2 = 1$, and the i-th entry of θ_1 and the j-th entry of $Q\theta_2$ are non-zero. If

$$Q_{ij}\text{sign}(\theta_1^*)_i\text{sign}(Q\theta_2^*) > 0,$$

then we have

$$\rho \ge \sqrt{(1-\sqrt{1-Q_{ij}^2})/2}. \tag{42}$$

Thus we will have chance to separate θ_1 and θ_2 successfully. It is easy to see that M is lower bounded by $\theta_1^T Q\theta_2$. Large $\theta_1^T Q\theta_2$ leads to larger M, but also leads to larger ρ, which is better for separating θ_1 and θ_2. The proof of above bound of ρ is given in appendix J.1.

In general, it is difficult for us to derive a lower bound of ρ like 42. Instead, we can derive the following sufficient condition in terms of M:

Theorem 8 *If $M \le \frac{1}{8\sqrt{s_1 s_2}}$, then for problem* (41) *with high probability*

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

When $s_1 = s_2 = 1$, this condition $M \le \frac{1}{8\sqrt{s_1 s_2}}$ is much stronger than (42), because every entry of Q has to be smaller than $1/8$;

D.2 Morphological Component Analysis Using k-support Norm

k-support norm [2] is another way to induce sparse solution instead of L1 norm. Recent works [2, 12] have shown that k-support norm has better statistical guarantee than L1 norm. For arbitrary $\theta \in \mathbb{R}^p$, its k-support norm $\|\theta\|_k^{sp}$ is defined as

$$\|\theta\|_k^{sp} = \inf\left\{\sum_{I\in\mathcal{G}_k} \|u_I\|_2 : \text{supp}(u_I) \subseteq I, \sum_{I\in\mathcal{G}_k} u_I = \theta\right\}.$$

For the superposition of an s_1 sparse vector and an s_2 sparse vector, the best choice is to use s_1-support norm and s_2-support norm. The new problem is

$$\min_{\theta_1,\theta_2} \|y - X(\theta_1+\theta_2)\|_2^2 \quad s.t. \quad \|\theta_1\|_{s_1}^{sp} \le \|\theta_1^*\|_{s_1}^{sp}, \|Q\theta_2\|_{s_2}^{sp} \le \|Q\theta_2^*\|_{s_2}^{sp}. \tag{43}$$

Denote $\sigma_{s_1,s_2}(Q)$ as the set of all the largest singular values of Q's $s_1 \times s_2$ submatrices. Let $\sigma = \max \sigma_{s_1,s_2}(Q)$. In this case, we have the following sufficient condition and high probability error bound:

Theorem 9 *If $\sigma \le \frac{1}{4(1+\frac{\theta_{1max}}{\theta_{1min}})(1+\frac{\theta_{2max}}{\theta_{2min}})}$ where $\theta_{max} = \max_{i\in\text{supp}(\theta)} |\theta_i|$ and $\theta_{min} = \min_{i\in\text{supp}(\theta)} |\theta_i|$, then we have for problem* (43) *with high probability*

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

In the problem setting of theorem 9, both norms are not decomposable. Therefore we can not apply the framework of [31] for this problem.

D.3 Low-rank and Sparse Matrix Decomposition

To recover a sparse matrix and low-rank matrix from their sum [8, 11], one can use k-support norm [2] 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. If we have $k > 1$, the framework in [31] is not applicable, because k-support norm is not decomposable. When $k = 1$, the component norms simplify to $\|\cdot\|_1$ for sparsity.

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. We want to recover S^* and L^* using s-support norm and nuclear norm respectively, so that the estimator takes the form:

$$\min_{L,S} \sum_{i=1}^{n} (Y_i - \langle X_i, L + S \rangle)^2 \quad s.t. \quad |||L|||_* \leq |||L^*|||_*, \|S\|_s^{sp} \leq \|S^*\|_s^{sp}. \tag{44}$$

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

Theorem 10 *If there is a $\rho > 0$ for problem* (44)*, then with high probability*

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

Theorem 10 requires SC condition to hold. When will SC condition for (44) holds? Early work have shown that to successfully estimate both L and S, the low-rank matrix L should satisfy "incoherence" condition [8]. From example in Appendix J.5, we can recovery matrix L and S even incoherence condition does not hold.

E Additional Experiments

E.1 Recovery using k-support norm

In our last experiment, we test the impact of sparsity on the estimation error. In this experiment we solve problem (43), and let both s_1 and s_2 vary from 2 to 3. We set the matrix Q to be a $p \times p$ discrete cosine transformation (DCT) matrix [14]. We use different problem size with $p = 100$ and $p = 150$. Number of samples n varies from 30 to 70. For each n, we generate 20 pairs of X and ω. For each (X, w) pair, we get a solution $\hat{\theta}_1$ and $\hat{\theta}_2$. We take the average over all $\|\hat{\theta}_1 - \theta_1^*\|_2 + \|\hat{\theta}_2 - \theta_2^*\|_2$. The plot is shown in figure 4. From figure 4, we can see that the error curve increases as dimensionality and sparsity increases. If p is fixed, then lower sparsity implies better estimation result.

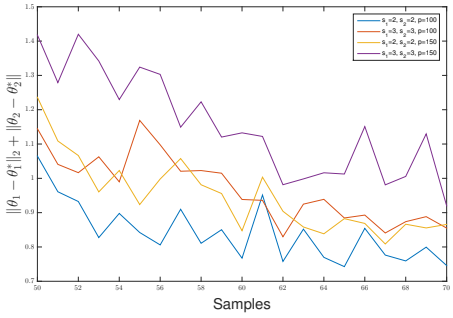


Figure 4: Effect of sparsity s_1, s_2 on estimation error. In all cases we use k-support norm instead of $l1$ norm.

F Proof of Theorem 1

Theorem 11 (Deterministic bound) *Assume that the RE condition in (6) is satisfied in $\mathcal{C}$ with parameter κ. Then, if $\kappa^2 > \gamma$, we have $\sum_{i=1}^k \|\Delta_i\|_2 \leqslant 2s_n(\gamma)$.*

Proof: By feasibility of θ^* and optimality of $\hat{\theta}$, we have

$$\|Y - X\hat{\theta}\|_2^2 \leq \|Y - X\theta^*\|_2^2 .$$

If $\hat{\boldsymbol{\theta}} = \sum_i \hat{\theta}_i$ is an optimum of (2), we have

$$\|Y - X\hat{\theta}\|_2^2 = \|X(\hat{\theta} - \theta^*)\|_2^2 - 2\omega^T X(\hat{\theta} - \theta^*) + \|\omega\|_2^2.$$

With $\Delta = \hat{\theta} - \theta^*$, $\Delta_i = \hat{\theta}_i - \theta_i^*$, we have

$$\|Y - X\hat{\theta}\|_2^2 - \|Y - X\theta^*\|_2^2 = \|X\Delta\|_2^2 - 2\omega^T X\Delta \leqslant 0. \tag{45}$$

For any $\Delta = \sum_{i=1}^k \Delta_i, \Delta_i \in \mathcal{C}_i$, if possible let $\sum_{i=1}^k \|\Delta_i\|_2 \geq 2s_n(\gamma)$. Then we have

$$\tfrac{1}{n}(\|X\Delta\|_2^2 - 2\omega^T X\Delta) \geq \left(\textstyle\sum_{i=1}^k \|\Delta_i\|_2\right)^2 (\kappa^2 - \gamma) > 0. \tag{46}$$

since $\kappa^2 > \gamma$. However, the inequality contradicts (45). Therefore $\sum_{i=1}^k \|\Delta_i\|_2 \leq 2s_n(\gamma)$. ∎

G Geometry of Structural Coherence

In this section, our goal is to characterize the geometric property of our SC condition. We start from a simple case when $k = 2$.

G.1 Proof of Lemma 2

Lemma 6 *If there exists a $\delta < 1$ such that $-\langle x, y\rangle \leq \delta\|x\|_2\|y\|_2$, then*

$$\|x + y\|_2 \geq \sqrt{\frac{1-\delta}{2}}(\|x\|_2 + \|y\|_2) . \tag{47}$$

Proof: We know from [19] that

$$\|x + y\|_2^2 \geq (1 - \delta)(\|x\|_2^2 + \|y\|_2^2)$$

and

$$(\|x\|_2 + \|y\|_2)^2 \leq 2(\|x\|_2^2 + \|y\|_2^2)$$

Combine them and we will get the conclusion. ∎

G.2 Proof of Theorem 3

Theorem 12 (Structural Coherence (SC) Condition) *Let $\delta := \max_i \delta_i$ with δ_i as defined in (14). If $\delta < 1$, 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.,*

$$\left\|\sum_{i=1}^k \Delta_i\right\|_2 \geq \rho \sum_{i=1}^k \|\Delta_i\|_2 . \tag{48}$$

Proof: We have by lemma (2)

$$\|\textstyle\sum_i \Delta_i\|_2 \geq \sqrt{\frac{1-\delta}{2}} \left(\|\Delta_{i'}\|_2 + \left\|\sum_{j\neq i'} \Delta_k\right\|_2\right) .$$

Sum over all possible combinations, we get

$$k\,\|\textstyle\sum_i \Delta_i\|_2 \geq \sqrt{\frac{1-\delta}{2}} \sum_{i'} \left(\|\Delta_{i'}\|_2 + \left\|\sum_{j\neq i'} \Delta_k\right\|_2\right) \geq \sqrt{\frac{1-\delta}{2}} \sum_{i'} \|\Delta_{i'}\|_2.$$

Therefore

$$\left\|\textstyle\sum_{i=1}^k \Delta_i\right\|_2 \geq \frac{1}{k}\sqrt{\frac{1-\delta}{2}} \sum_{i=1}^k \|\Delta_i\|_2.$$

∎

H Restricted Eigenvalue Condition

H.1 Proof of Lemma 2

Lemma 7 *Let sets $\mathcal{C}$ 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:*

$$Q_{\rho\xi}(\mathcal{H}; Z) \geq Q_{\xi}(\mathcal{A}; Z). \tag{49}$$

Proof: By definition, for any $u \in \mathcal{H}$, we can find $u_i \in \mathcal{C}_i, i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} u_i = u$. Then

$$|\langle Z, u\rangle| = \left\|\sum_{i=1}^{k} u_i\right\|_2 \left|\left\langle Z, \frac{u}{\|\sum_{i=1}^{k} u_i\|_2}\right\rangle\right| \geq \rho \left|\left\langle Z, \frac{u}{\|\sum_{i=1}^{k} u_i\|_2}\right\rangle\right|.$$

Let $v = \frac{u}{\|\sum_{i=1}^{k} u_i\|_2}$, from definition we know that $v \in A$. Hence we also have

$$P(|\langle Z, u\rangle| \geq \rho\xi) = P\left(\tfrac{1}{\rho}|\langle Z, u\rangle| \geq \xi\right) \geq P(|\langle Z, v\rangle| \geq \xi).$$

Therefore taking the infimum over all $v \in \mathcal{A}$ and then all $u \in \mathcal{C}$, the conclusion holds. ∎

H.2 Proof of Theorem 4

Theorem 13 (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*

$$\inf_{u \in \mathcal{H}} \|Xu\|_2 \geq c_1 \rho \sqrt{n} - c_2 w(\mathcal{H}) - c_3 \rho t \tag{50}$$

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

Proof: Let two sets $\mathcal{C}$ and $\mathcal{A}$ be as defined previously. From Lemma (1) and lemma (2) we know that for any $\xi > 0$, with probability at least $1 - e^{-t^2/2}$

$$\inf_{u \in \mathcal{H}} \|Xu\|_2 \geq \rho\xi\sqrt{n} Q_{2\xi}(\mathcal{A}; Z) - 2W(\mathcal{H}; Z) - \rho\xi t. \tag{51}$$

We use the "Bowling scheme" in [27], let $\mathbf{v}$ be any vector in A, by Paley-Zygmund inequality [6], one can get

$$P(|\langle x, v\rangle| \geq 2\xi) \geq \frac{[E|\langle x, v\rangle| - 2\xi]_+^2}{E|\langle x, v\rangle|^2} \geq \frac{(\alpha - 2\xi)^2}{4\sigma_x^2}. \tag{52}$$

From the proof of [27, Theorem 6.3], empirical width can be bounded by

$$W(\mathcal{H}; Z) \leq L\sigma_x w(\mathcal{H}) \tag{53}$$

Select $\xi = \alpha/6$, combine (51), (52), (53) to discover that:

$$\inf_{u \in \mathcal{C}} \|Xu\|_2 \geq \frac{1}{9}\rho\alpha^3\sigma_x^{-2}\sqrt{n} - 2L\sigma_x w(\mathcal{H}) - \rho\frac{\alpha}{6}\tau$$

which completes the proof. ∎

From the conclusion above, the right hand side contains three parts. The first part is a constant times the square root of sample size, and the second part is a measure of the complexity of error sets. Therefore, when the number of samples is large enough or the error set has low complexity, the right terms will be larger than zero.

H.3 Proof of Proposition 5

Proposition 14 *If there is a matrix X such that condition* (6) *holds for $\Delta_i \in \mathcal{C}_i$, then SC (9) holds.*

Proof: If such ρ does not exist, then there are some $\Delta_i \in \mathcal{C}_i, i = 1, \ldots, k$ not all zero such that

$$\left\|\sum_{i=1}^{k} \Delta_i\right\|_2 = 0 \Rightarrow \sum_{i=1}^{k} \Delta_i = 0,$$

which implies $\|X\sum_{i=1}^{k} \Delta_i\|_2 = 0$ for every X. This is a contradiction. ∎

I Error Bound

I.1 Proof of Lemma 3

Lemma 8 *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 $s_n(\gamma) \le c\frac{w(\bar{\mathcal{H}})}{\gamma\sqrt{n}}$. 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.*

Proof: First notice that

$$\omega^T X\Delta = \|\omega\|_2 . \frac{\omega X\Delta}{\|\omega\|_2}.$$

We can first bound $\frac{\omega X\Delta}{\|\omega\|_2}$ then bound $\|\omega\|_2$.

(a). Bound $\frac{\omega X\Delta}{\|\omega\|_2}$: Note that $\frac{\omega X\Delta}{\|\omega\|_2}$ is not centered. In the first step we center it using

$$\frac{1}{\|\omega\|_2}\epsilon_i\omega_i x_i^T\Delta,$$

where ϵ_i is a Radmacher random variable, its probability of being $+1$ and -1 are both half. $x_i \in R^p$ is the i-th row of X. By assumption we know different x_i are independent and have same distribution. Here we fix ω. By proposition 5.10 in [28], the following bound holds:

$$P\left(\left|\frac{\omega X\Delta}{\|\omega\|_2}\right| \ge t\right) \le e.\exp\left(-\frac{c_1 t^2}{\sigma_x^2\|\Delta\|_2}\right). \tag{54}$$

Through simple transform we know that

$$P\left(\left|\left\langle \frac{\omega X}{\|\omega\|_2}, u - v\right\rangle\right| \ge t\right) \le e.e^{-c_1 t^2/(\sigma_x^2\|u-v\|_2^2)}$$

for any $u, v \in \mathbb{R}^p$. Then use [26, Theorem 2.2.27]

$$P\left(\sup_{\Delta\in\bar{\mathcal{H}}}\left|\frac{1}{\|\omega\|_2}\sum_i \epsilon_i\omega_i x_i^T\Delta\right| \ge c_2 w(\bar{\mathcal{H}}) + 2c_3 t\right) \le c_4\exp\left(-\frac{4c_1 t^2}{\sigma_x^2}\right)$$

(b). Bound ω: We first notice that $\|\omega\|_2^2$ is a sub-exponential random variable. Therefore if the sub-gaussian norm of ω is $\|\omega\|_{\phi_2}$, then for each entry ω_i the sub-expoential norm $\|\omega_i^2\|_{\phi_1} \le 2\|\omega\|_{\phi_2}^2$. Through definition we reach:

$$\mathbb{E}\|\omega\|_2 \le \sqrt{\mathbb{E}\|\omega\|_2^2} \le \sigma_\omega\sqrt{2n}.$$

Applying proposition 5.16 in [28] to $\|\omega\|_2^2$, we obtain

$$P(|\|\omega\|_2^2 - \mathbb{E}\|\omega\|_2^2| \ge t) \le 2\exp\left[-c_5\min\left(\frac{t^2}{4\sigma_\omega^4}, \frac{t}{2\sigma_\omega^2}\right)\right].$$

Replace t with $\sigma_\omega^2 n$ gives

$$P(\|\omega\|_2 \ge 2\sigma_x\sqrt{n}) \le 2\exp(-c_5 n).$$

Combine (a) and (b): First we have

$$\begin{aligned}
&P\left(\sup_{\Delta\in\bar{\mathcal{H}}}\frac{1}{\sqrt{n}}\left|\sum_i \epsilon_i\omega_i x_i^T\Delta\right| \le 2c_2\sigma_x w(\bar{\mathcal{H}}) + 4c_3\sigma_x t\right)\\
\ge& P\left(\sup_{\Delta\in\bar{\mathcal{H}}}\left|\frac{1}{\|\omega\|_2}\sum_i \epsilon_i\omega_i x_i^T\Delta\right| \ge c_2 w(\bar{\mathcal{H}}) + 2c_3 t \wedge \frac{\|\omega\|_2}{\sqrt{n}} \ge 2\sigma_x\right)\\
\ge& P\left(\sup_{\Delta\in\bar{\mathcal{H}}}\left|\frac{1}{\|\omega\|_2}\sum_i \epsilon_i\omega_i x_i^T\Delta\right| \ge c_2 w(\bar{\mathcal{H}}) + 2c_3 t|\omega\right) P\left(\frac{\|\omega\|_2}{\sqrt{n}} \ge 2\sigma_x\right)\\
\ge& 1 - 2\exp(-c_5 n) - c_4\exp\left(-\frac{4c_1 t^2}{\sigma_x^2}\right)
\end{aligned}$$

Then choose $t = \frac{c_2}{2c_3} w(\bar{\mathcal{H}})$,

$$P\left(\sup_{\Delta\in\bar{\mathcal{H}}} \frac{1}{\sqrt{n}} \left|\sum_i \epsilon_i \omega_i x_i^T \Delta\right| \le 4c_2\sigma_\omega w(\bar{\mathcal{H}})\right) \ge 1 - c_4.\exp\left(-\frac{c_1 c_2^2 w(\bar{\mathcal{H}})}{c_3\sigma_x^2}\right) - 2\exp(-c_5 n)$$

Now we have bounded the symmetrized $\omega^T X\Delta$, then $\omega^T X\Delta$ can be bounded using symmetrization of probability [17]:

$$\begin{aligned}
&P\left(\sup_{\Delta\in\bar{\mathcal{H}}} \frac{1}{\sqrt{n}} |\omega^T X\Delta - E\omega^T X\Delta| > 16c_5\sigma_\omega w(\bar{\mathcal{H}})\right) \\
\le& 4P\left(\sup_{\Delta\in\bar{\mathcal{H}}} \frac{1}{\sqrt{n}} \left|\sum_i \epsilon_i \omega_i x_i^T \Delta\right| > 4c_5\sigma_\omega w(\bar{\mathcal{H}})\right) \le 4c_4.\exp\left(-\frac{c_1 c_2^2 w(\bar{\mathcal{H}})}{c_3\sigma_x^2}\right) + 8\exp(-c_5 n).
\end{aligned}$$

Because ω has zero mean, there the above inequality give us:

$$\sup_{\Delta\in\bar{\mathcal{H}}} \frac{|\omega^T X\Delta|}{\sqrt{n}} \le 16c_5\sigma_\omega w(\bar{\mathcal{H}})$$

with high probability. Hence by definition

$$s_n(\gamma) \le \frac{16c_5\sigma_\omega w(\bar{\mathcal{H}})}{\gamma\sqrt{n}}$$

with probability at least $1 - 4c_4.\exp\left(-\frac{c_1 c_2^2 w(\bar{\mathcal{H}})}{c_3\sigma_x^2}\right) - 8\exp(-c_5 n)$. ∎

Now as we have the high probability bound of both κ and $s_n(\gamma)$, we can derive our error bound for random case.

I.2 Proof of Lemma 4

Lemma 9 (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 $w(\bar{\mathcal{H}}) = O\left(\max_i w(\mathcal{C}_i \cap B_p) + \sqrt{\log k}\right)$.*

Proof: By definition and the fact that the Gaussian width of convex hull of sets is equal to the Gaussian width of their union [10]

$$w(\mathcal{H}) = \mathbb{E}\sup_{u\in\mathcal{H}} \langle u, g\rangle = \mathbb{E}\max_i \sup_{u_i\in\mathcal{C}_i\cap S_{p-1}} \langle u_i, g\rangle$$

By concentration inequality for Lipschitz functions [17], for each $i = 1, \ldots, k$

$$P(\sup_{u_i\in\mathcal{C}_i\cap S_{p-1}} \langle u_i, g\rangle \ge \mathbb{E}\sup_{u_i\in\mathcal{C}_i\cap S_{p-1}} \langle u_i, g\rangle + r) \le \exp(-r^2/2).$$

Then denote $D_i = \sup_{u_i\in\mathcal{C}_i\cap S_{p-1}} \langle u_i, g\rangle$, we have

$$\begin{aligned}
w(\mathcal{H}) =& \mathbb{E}\max_i D_i \\
&\le \max_i \mathbb{E}D_i + \delta + \sum_{i=1}^k \int_\delta^\infty P\left(D_i \ge \max_i \mathbb{E}D_i + r\right) dr \\
&\le \max_i \mathbb{E}D_i + \delta + k\int_\delta^\infty \exp(-r^2/2) dr \\
&\le \max_i \mathbb{E}D_i + \delta + \eta' k\exp(-\delta^2/2).
\end{aligned}$$

Let $\delta = \sqrt{\log k}$, we get

$$w(\mathcal{H}) \le \max_i \mathbb{E}D_i + \sqrt{\log k} + \eta = O\left(\max_i w(\mathcal{C}_i \cap S_{p-1}) + \sqrt{\log k}\right).$$

the conclusion holds. The conclusion for $w(\bar{\mathcal{H}})$ can be proved the same as above. ∎

I.3 Proof of Theorem 6

Theorem 15 *For estimator (3), 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 a 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 with high probability,*

$$\sum_{i=1}^{k} \|\hat{\theta}_i - \theta_i^*\|_2 \leq C \frac{\max_i w(\mathcal{C}_i \cap B_p) + \sqrt{\log k}}{\rho^2 \sqrt{n}}, \tag{55}$$

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

Proof: Firstly, we choose

$$t = \frac{1}{3}\alpha^2 \sigma_x^{-2} \sqrt{n} - 6L\sigma_x \rho^{-1} \alpha^{-1} w(\mathcal{H}).$$

From theorem 4, RE condition holds for

$$\kappa \geq \frac{1}{2}\left(\frac{1}{9}\rho\alpha^3\sigma_x^{-2} - 2L\sigma_x \frac{w(\mathcal{H})}{\sqrt{n}}\right)$$

with probability at least

$$1 - \exp\left(-(\frac{1}{3}\alpha^2\sigma_x^{-2}\sqrt{n} - 6L\sigma_x\rho^{-1}\alpha^{-1}w(\mathcal{H}))^2/2\right).$$

Next we choose $\gamma = \frac{\rho^2\alpha^6}{1296\sigma_x^4}$, and let $s_n(\gamma)$ be defined as above. Thus from theorem (1) and our discussion above, if

$$\frac{1}{4}\left(\frac{1}{9}\rho\alpha^3\sigma_x^{-2} - 2L\sigma_x\frac{w(\mathcal{H})}{\sqrt{n}}\right)^2 > \frac{\rho^2\alpha^6}{1296\sigma_x^4} \Rightarrow n > c_1 w^2(\mathcal{H}),$$

for some constant $c_1 > 0$, then $\kappa > 2\gamma$. Using theorem 1, we have

$$\sum_{i=1}^{k} \|\Delta_i\|_2 \leq c_2 \frac{w(\bar{\mathcal{H}})}{\sqrt{n}}.$$

with probability at least

$$1 - c_3\exp(-c_4 w(\bar{\mathcal{H}})) - c_5\exp(-c_6 n) - \exp(-(c_7\sqrt{n} - c_8 w(\mathcal{H}))^2),$$

which completes the proof. ∎

J Examples

J.1 Structural Coherence For 1-sparse + 1-sparse MCA

Proposition 16 *Suppose both vector θ_1 and vector $Q\theta_2$ are one sparse, and the i-th entry of θ_1 and the j-th entry of $Q\theta_2$ are non-zero. If*

$$Q_{ij}\text{sign}(\theta_1^*)_i\text{sign}(Q\theta_2^*) > 0,$$

then we have

$$\rho \geq \sqrt{(1 - \sqrt{1 - Q_{ij}^2})/2}. \tag{56}$$

Proof: Denote Δ_{-i} as a vector whose ith entry is 0, and other entries are equal to those of Δ. Suppose both θ_1 and $Q\theta_2$ are 1-sparse vectors, and the ith entry of θ_1 jth entry of $Q\theta_2$ are nonzero. Then error vector Δ_1 and Δ_2 satisfy the following inequalities:

$$-\langle\text{sign}(\theta_{1i}), \Delta_{1i}\rangle \geq \|\Delta_{1-i}\|_1, -\langle\text{sign}(\theta_{2j}), \Delta_{2j}\rangle \geq \|\Delta_{2-j}\|_1.$$

Therefore

$$\frac{-\langle \text{sign}(\theta_{1i}), \Delta_{1i}\rangle}{\|\Delta_1\|_2} \geq \frac{1}{\sqrt{1+\frac{\|\Delta_{1-i}\|_2^2}{\Delta_{1i}^2}}} \geq \frac{1}{\sqrt{1+\frac{\|\Delta_{1-i}\|_2^2}{\|\Delta_{1-i}\|_1^2}}} \geq \frac{1}{\sqrt{2}}.$$

The same holds for $Q\Delta_2$. Then when $Q_{ij}\text{sign}(\theta_{1i})\text{sign}(Q\theta_{2j}) > 0$ we have from geometry that

$$-\langle \Delta_1, \Delta_2\rangle \leq -\cos(2\arccos(\frac{1}{\sqrt{2}}) + \arccos(Q_{ij}\text{sign}(\theta_{1i})\text{sign}(Q\theta_{2j}))) \leq \sqrt{1-Q_{ij}^2}$$

Therefore $\rho \geq \sqrt{\frac{1-\sqrt{1-Q_{ij}^2}}{2}}$ by proposition 2. ∎

J.2 Proof of Theorem 8

Theorem 17 *If $M \leq \frac{1}{8\sqrt{s_1 s_2}}$, then for problem* (41) *with high probability*

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

Proof: Let $\Delta_i = \theta_i - \theta_i^*$ for $i = 1, 2$, then we have

$$\begin{aligned}\langle \Delta_1, \Delta_2\rangle = \langle \Delta_1, Q^T Q\Delta_2\rangle &\leq \max_{ij}|Q_{ij}|\|\Delta_1\|_1\|Q\Delta_2\|_1 \\ &\leq M(\|P_\Omega \Delta_1\|_1 + \|P_{\Omega_1^c}\Delta_1\|_1)(\|P_{\Omega_2}Q\Delta_2\|_1 + \|P_{\Omega_2^c}Q\Delta_2\|_1) \\ &\leq 4M\|P_\Omega\Delta_1\|_1\|P_{\Omega_2}Q\Delta_2\|_1 \leq 4M\sqrt{s_1 s_2}\|\Delta_1\|_2\|Q\Delta_2\|_2\end{aligned}$$

Let $4M\sqrt{s_1 s_2} \leq \frac{1}{2}$, we get $M \leq \frac{1}{8\sqrt{s_1 s_2}}$ and $\rho \geq \frac{1}{2}$.

From the result of [10], we know that the Gaussian width for the error cone of a s-sparse vector is $O(\sqrt{s\log(\frac{p}{s})})$. Therefore by theorem 6, if $n \geq C\max_{s\in\{s_1,s_2\}} s\log(\frac{p}{s})$ for some $C > 0$, then

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

∎

J.3 Proof of Theorem 9

Theorem 18 *If $\sigma \leq \frac{1}{4(1+\frac{\theta_{1max}}{\theta_{1min}})(1+\frac{\theta_{2max}}{\theta_{2min}})}$ where $\theta_{max} = \max_{i\in\text{supp}(\theta)}|\theta_i|$ and $\theta_{min} = \min_{i\in\text{supp}(\theta)}|\theta_i|$, then we have for problem* (43) *with high probability*

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

Proof: We first characterize the interaction between cones:

$$\langle \Delta_1, \Delta_2\rangle = \langle \Delta_1, Q^T Q\Delta_2\rangle.$$

because k-support norm is an atomic norm, and its atomic set is all unit k-sparse vectors. Therefore we can decompose Δ_1 into combination of unit s_1-sparse vectors $\Delta_1 = \sum_i \alpha_i u_i$ and $Q\Delta_2$ into combination of unit s_2-sparse vectors $Q\Delta_2 = \sum_j \beta_j v_j$. Then

$$\langle \Delta_1, Q^T Q\Delta_2\rangle \leq \sum_i |\alpha_i| \sum_j |\beta_j| \max_{ij}|u_i^T Q^T v_j| \leq \sigma\|\Delta_1\|_{s_1}^{sp}\|\Delta_2\|_{s_2}^{sp}.$$

By Theorem 9 in [13], we have

$$\|\Delta_1\|_{s_1}^{sp} \leq \sqrt{2}(1+\frac{\theta_{1max}}{\theta_{1min}})\|\Delta_1\|_2 \quad \text{and} \quad \|\Delta_2\|_{s_2}^{sp} \leq \sqrt{2}(1+\frac{\theta_{2max}}{\theta_{2min}})\|\Delta_1\|_2.$$

Therefore

$$\langle \Delta_1, \Delta_2 \rangle \leq 2\sigma(1 + \frac{\theta_{1max}}{\theta_{1min}})(1 + \frac{\theta_{2max}}{\theta_{2min}})\|\Delta_1\|_2\|\Delta_2\|_2.$$

Let $2\sigma(1 + \frac{\theta_{1max}}{\theta_{1min}})(1 + \frac{\theta_{2max}}{\theta_{2min}}) \leq \frac{1}{2}$, then $\sigma \leq \frac{1}{4(1+\frac{\theta_{1max}}{\theta_{1min}})(1+\frac{\theta_{2max}}{\theta_{2min}})}$ and $\rho \geq \frac{1}{2}$.

From [13], when we set k to be the sparsity s, the corresponding Gaussian width of tangent cone is $O(s\log(p-s))$, therefore plus this result in Theorem 6 we get

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

■

J.4 Proof of Theorem 10

Theorem 19 *If there is a $\rho > 0$ for problem* (44)*, then with high probability*

$$\|L - L^*\|_2 + \|S - S^*\|_2 = O\left(\max\left\{\sqrt{\frac{s\log(d_1d_2 - s)}{n}}, \sqrt{\frac{r(d_1 + d_2 - r)}{n}}\right\}\right).$$

Proof: From [10], we know that for error cone of $d_1 \times d_2$ rank-r matrix, its Gaussian width is $O(r(d_1 + d_2 - r))$. Therefore if $\rho > 0$, then by applying theorem 6, the error bound is

$$\|L - L^*\|_2 + \|S - S^*\|_2 = O\left(\max\left\{\sqrt{\frac{s\log(d_1d_2 - s)}{n}}, \sqrt{\frac{r(d_1 + d_2 - r)}{n}}\right\}\right).$$

■

J.5 Additional Example for Low-rank ans Sparse Matrix Decomposition

Example 1 *Suppose noise $\omega = 0$,*

$$S_0 = L_0 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \tag{57}$$

and $M = S_0 + L_0$, then the SC condition of problem (44) holds.

Proof: Suppose the singular value decomposition of L_0^* is $U\Sigma V^T$. Denote

$$\mathcal{C}_S' = cl\{\Delta | - \langle sign(S_0'), \Delta\rangle \geq \|P_{\Omega^c}(\Delta)\|_1\},\ \mathcal{C}_L' = cl\{\Delta | - \langle UV^T, \Delta\rangle \geq \|P_{T^\perp}(\Delta)\|_*\}.$$

From [19] we know that SC is equivalent to $-\mathcal{C}_S' \bigcap \mathcal{C}_L' = \{0\}$. To prove $-\mathcal{C}_S' \bigcap \mathcal{C}_L' = \{0\}$, we need the following inequalities:

$$\langle sign(S_0'), \Delta\rangle \geq \|P_{\Omega^c}(\Delta)\|_1,\ -\langle UV^T, \Delta\rangle \geq \|P_{T^\perp}(\Delta)\|_*.$$

has unique solution 0.

It is easy to notice that $\langle sign(S_0'), \Delta\rangle = \langle UV^T, \Delta\rangle = \Delta_{11}$. As the value of norms is non-negative, we have $\Delta_{11} \geq 0$ and $-\Delta_{11} \geq 0$. Therefore $\Delta_{11} = 0$. Besides,

$$\|P_{\Omega^c}(\Delta)\|_1 = \textstyle\sum_{(i,j)\neq(1,1)} |\Delta_{ij}| \leq 0,$$

which leads to $\Delta_{ij} = 0$ for $(i,j) \neq (1,1)$.

Finally, $\Delta = 0$ and the conclusion holds. ■

People tend to think that we cannot obtain the correct decomposition in this situation. Note that the cone $\mathcal{C}_S$ is centered at one point, and the cone $\mathcal{C}_L$ contains $\mathcal{C}_S$ but their surface contacts only at the origin. Therefore the reflection of one cone will touch the other cone only at the origin. As a result, for $M = M_0 = S_0 + L_0$, i.e., SC condition holds.

K Noiseless Case: Comparing Estimators

In this section we try to explore the structures that are different between problem (2) and problem (36) and the structures that they share.

K.1 Proof of Lemma 5

Lemma 10 *For a given set $\{\lambda_i\}$, the infimal convolution norm ball Ω_R is the convex hull of $\bigcup_{i=1}^k \frac{1}{\lambda_i}\Omega_R^i$, i.e., $\Omega_R = \mathrm{conv}(\bigcup_{i=1}^k \frac{1}{\lambda_i}\Omega_R^i)$.*

Proof: If $\theta \in \Omega$, then from definition there are $\sum_{i=1}^k \theta_i = \theta$ such that $R(\theta) = \sum_{i=1}^k \lambda_i R_i(\theta_i)$. Without loss of generalization, suppose $\theta_i \neq 0$ for each i, then we have the following decomposition:

$$\sum_{i=1}^k \frac{\lambda_i R_i(\theta_i)}{R(\theta)} \frac{R(\theta)}{\lambda_i R_i(\theta_i)} \theta_i = \theta. \tag{58}$$

It is easy to know that $\sum_{i=1}^k \frac{\lambda_i R_i(\theta_i)}{R(\theta)} = 1$ and $R_i(\frac{R(\theta)}{\lambda_i R_i(\theta_i)}\theta_i) = \frac{1}{\lambda} R(\theta) \leq \frac{1}{\lambda_i}$. Therefore $\frac{R(\theta)}{\lambda_i R_i(\theta_i)}\theta_i \in \frac{1}{\lambda}\Omega_R^i$, and $\theta \in conv(\bigcup_{i=1}^k \frac{1}{\lambda_i}\Omega_R^i)$.

If $\theta \in conv(\bigcup_{i=1}^k \frac{1}{\lambda_i}\Omega_R^i)$, then we can find $\theta_i \in \frac{1}{\lambda_i}\Omega_R^i$ and $c_i > 0, \sum_{i=1}^k c_i = 1$ such that

$$\sum_{i=1}^k c_i \theta_i = \theta.$$

Then

$$R(\theta) \leq \sum_{i=1}^k \lambda_i R_i(c_i \theta_i) = \sum_{i=1}^k c_i \lambda_i R_i(\theta_i) \leq 1.$$

Therefore $\theta \in \Omega_R$ which completes the proof. ∎

K.2 Proof of theorem 7

Theorem 20 *Given $\hat{\theta}_1, \ldots, \hat{\theta}_k$ and define*

$$\mathcal{C}_0 = \left\{ \sum_{\theta_i \neq 0} \left(\frac{c_i'}{c_i} - 1 \right) \theta_i \mid c_i' \geq 0, \sum_{i=1}^k c_i' = 1 \right\}. \tag{59}$$

Suppose $dim(span\{\theta_i\}) = k$, then there exist $\lambda_1, \ldots, \lambda_k$ such that $\sum_{i=1}^k \hat{\theta}_i = \theta$ are unique solutions of (37) if and only if there are $c_1, \ldots, c_k$ with $c_i \geq 0$ and $\sum_{i=1}^k c_i = 1$ such that for the corresponding error cone $\mathcal{C}_i$ of $\hat{\theta}_i$ and C_0 defined above, $-\mathcal{C}_i \cap \sum_{j \neq i} \mathcal{C}_j = \{0\}$, for $i = 0, 1, \ldots, k$.

Proof: Before proofing the main result we need the following lemma, it is proved in appendix K.3.

Lemma 11 *For fixed $\lambda_1, \ldots, \lambda_k$, suppose $\sum_{i=1}^k \theta_i = \theta$ is a solution of decomposition (37) under this set of $\{\lambda_i\}$, and $dim(span\{\theta_i\}) = k$. Let $\mathcal{C}_i, i = 1, 2, \ldots, k$ be the corresponding error cones of $\{\theta_i\}$, $c_i = \lambda_i R_i(\theta_i)/R(\theta)$ and $\mathcal{C}_0$ be as defined in* (59)*. The decomposition (37) for θ is unique if and only if for any $i = 0, 1, \ldots, k$,*

$$-\mathcal{C}_i \cap \sum_{j \neq i} \mathcal{C}_j = \{0\}. \tag{60}$$

We come to the main result and the necessity is obvious from lemma 11;

Without loss of generality, suppose $c_i \geq 0$. If such $c_1, \ldots, c_k$ exist for $\hat{\theta}_1, \ldots, \hat{\theta}_k$, let $\lambda_i = \frac{c_i}{R_i(\hat{\theta}_i)}$. Suppose $\theta_1, \ldots, \theta_k$ is a set of optimal solution under the λ defined above. From 58 we can write the decomposition as $\theta = \sum_{i=1}^k c_i' \theta_i'$ where c_i' is a coefficient of convex combination, $c_i'\theta_i' = \theta_i$ and $\lambda_i R_i(\theta_i') = R(\theta)$ under coefficient λ_i.

If $\theta_i \neq \hat{\theta}_i$ for some i, then as $\sum_{i=1}^k \lambda_i R_i(\hat{\theta}_i) = \sum_{i=1}^k c_i = 1$, we have

$$\lambda_i R_i(\theta_i) \leq 1 \Rightarrow R_i(c_i\theta_i) \leq R_i(\hat{\theta}_i).$$

Therefore $c_i\theta_i - \hat{\theta}_i \in \mathcal{C}_i$ by definition. We also have

$$\sum_{i=1}^k \hat{\theta}_i + \sum_{i=1}^k (\frac{c_i'}{c_i} - 1)\hat{\theta}_i + \sum_{i=1}^k \frac{c_i'}{c_i}(c_i\theta_i - \theta_i) = \theta \Rightarrow \frac{c_i'}{c_i}(c_i\theta_i - \theta_i) = -\sum_{i=1}^k (\frac{c_i'}{c_i} - 1)\hat{\theta}_i$$

which is contradict to our condition. Therefore $\hat{\theta}_i$ is a solution. Uniqueness is a direct conclusion of lemma 11. ■

Note that in this proof we set $\lambda_i = \frac{c_i}{R_i(\theta_i)}$. This is also a general way to choose the parameter $\{\lambda_i\}$.

K.3 Proof of Lemma 11

Proof: Without loss of generality, assume $\theta_i \neq 0$.According to (58), let

$$c_i = \frac{\lambda_i R_i(\theta_i)}{R(\theta)}, and\ \ \theta_i' = \frac{1}{c_i}\theta_i.$$

Suppose θ_i is a unique decomposition, for any $\Delta_i \in \mathcal{C}_i$, if $\sum_{i=1}^k c_i'\lambda_i R_i(\theta_i' + \Delta_i) \leq R(\theta)$ and $\sum_{i=1}^k c_i'(\theta_i' + \Delta_i) = \theta$ for some $c_i' \geq 0, \sum_{i=1}^k c_i' = 1$. Therefore we obtain the following decomposition of θ:

$$\theta = \sum_{i=1}^k c_i\theta_i' + \sum_{i=1}^k (c_i' - c_i)\theta_i + \sum_{i=1}^k c_i'\Delta_i.$$

It is obvious from observation that $\sum_{i=1}^k (c_i' - c_i)\theta_i + \sum_{i=1}^k c_i'\Delta_i = 0$ and $\sum_{i=1}^k (c_i' - c_i)\theta_i \in \mathcal{C}_0$, $\sum_{i=1}^k c_i'\Delta_i \in \sum_{i=1}^k \mathcal{C}_i$.

By minimal of $R(\theta)$, $\sum_{i=1}^k c_i'\lambda_i R_i(\theta_i' + \Delta_i) = R(\theta)$. By our assumption, such decomposition of θ is unique, thus

$$\sum_{i=1}^k (c_i' - c_i)\theta_i = \sum_{i=1}^k c_i'\Delta_i = 0.$$

which implies $-\mathcal{C}_0 \cap \sum_{i=1}^k \mathcal{C}_i = \{0\}$. Uniqueness also give that $c_i'\Delta_i = 0$ for each $i = 1, 2, \ldots, k$. Therefore $-\mathcal{C}_i \cap \sum_{j\neq i} \mathcal{C}_j = \{0\}$.

If θ_i is not a unique decomposition then there are some $\Delta_i \neq 0$ such that $\sum_{i=1}^k \Delta_i = 0$ and

$$\sum_{i=1}^k \lambda_i R_i(\theta_i + \Delta_i) = R(\theta).$$

Let

$$c_i'' = \frac{\lambda_i R_i(\theta_i + \Delta_i)}{R(\theta)}, and\ \ \theta_i'' = \frac{1}{c_i''}(\theta_i + \Delta_i),$$

we have

$$\lambda_i R_i(\theta_i'') = R(\theta),$$

and hence $\theta_i'' - \theta_i' \in \mathcal{C}_i$ for $\lambda_i R_i(\theta_i') = R(\theta)$. Unfold θ_i' and θ_i'' gives

$$c_i''(\theta_i'' - \theta_i') = c_i''(\frac{1}{c_i''} - \frac{1}{c_i'})\theta_i + \Delta_i.$$

Sum over all i, we get

$$\sum_{i=1}^k c_i''(\theta_i'' - \theta_i') = \sum_{i=1}^k c_i''(\frac{1}{c_i''} - \frac{1}{c_i'})\theta_i,$$

which is contradict to our assumption that $-\mathcal{C}_0 \cap \sum_{i=1}^k \mathcal{C}_i = \{0\}$. Therefore the conclusion holds. ■

Supplement References

[25] 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.

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

[27] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[28] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013.

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

[30] 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.

[31] 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.

[32] S. Chatterjee, S. Chen, and A. Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems 27*. 2014.

[33] S. Chen and A. Banerjee. Structured estimation with atomic norms: General bounds and applications. In *Advances in Neural Information Processing Systems 28*. 2015.

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

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

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

[37] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *A Series of Modern Surveys in Mathematics*. Springer Berlin, 2013.

[38] M. B. McCoy. A geometric analysis of convex demixing. *Ph.D. Thesis, Caltech*, 2013.

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

[40] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

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

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

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

[44] 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.

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

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