Chapter 9
Improved Sample Complexity in Sparse Subspace Clustering with Noisy and Missing Observations

In this chapter, we show the results of the new CoCoSSC algorithm. The content is organized as follows: The main results concerning CoCoSSC algorithm are shown in Sect. 9.1. Following Sect. 9.1, we show the full proofs in Sect. 9.2. In Sect. 9.3, we show the performance for CoCoSSC algorithm and some related algorithms numerically. Finally, we conclude this work with some future directions.

9.1 Main Results About CoCoSSC Algorithm

We introduce our main results by analyzing the performance of CoCoSSC under both the Gaussian noise model and the missing data model. Similar to [WX16], the quality of the computed self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ is assessed using a *subspace detection property (SDP)*:

Definition 9.1 (Subspace Detection Property (SDP), [WX16]) The self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ satisfies the *subspace detection property* if (1) for every $i \in [N]$, $\boldsymbol{c}_i$ is a non-zero vector; and (2) for every $i, j \in [N]$, $\boldsymbol{c}_{ij} \neq 0$ implies that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster.

Intuitively, the subspace detection property asserts that the self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ has *no false positives*, where every non-zero entry in $\{\boldsymbol{c}_i\}_{i=1}^n$ links two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to the same cluster. The first property in Definition 9.1 further rules out the trivial solution of $\boldsymbol{c}_i \equiv \boldsymbol{0}$.

Part of this chapter is in the paper titled "Improved Sample Complexity in Sparse Subspace Clustering with Noisy and Missing Observations" by Yining Wang, Bin Shi et al. (2018) presently under review for publication in AISTATS.

B. Shi, S. S. Iyengar, *Mathematical Theories of Machine Learning - Theory and Applications*, https://doi.org/10.1007/978-3-030-17076-9_9

The SDP stated in Definition 9.1 is, however, *not* sufficient for the success of a follow-up spectral clustering algorithm, or any clustering algorithm, as the “similarity graph” constructed by connecting every pairs of (i, j) with $\boldsymbol{c}_{ij} \neq 0$ might be poorly connected. Such “graph connectivity” is a well-known open problem in sparse subspace clustering [NH11] and remains largely unsolved except under strong assumptions [WWS16]. Nevertheless, in practical scenarios the SDP criterion correlates reasonably well with clustering performance [WX16, WWS15a] and therefore we choose to focus on the SDP success condition only.

9.1.1 The Non-Uniform Semi-Random Model

We adopt the following non-uniform semi-random model throughout the paper:

Definition 9.2 (Non-Uniform Semi-Random Model) Suppose $\boldsymbol{y}_i$ belongs to cluster $\mathcal{S}_\ell$ and let $\boldsymbol{y}_i = \mathbf{U}_\ell \boldsymbol{\alpha}_i$, where $\mathbf{U}_\ell \in \mathbb{R}^{n \times d_\ell}$ is an orthonormal basis of $\mathcal{U}_\ell$ and $\boldsymbol{\alpha}_i$ is a d_ℓ-dimensional vector with $\|\boldsymbol{\alpha}_i\|_2 = 1$. We assume that $\boldsymbol{\alpha}_i$ are i.i.d. distributed according to an unknown underlying distribution P_ℓ, and that the density p_ℓ associated with P_ℓ satisfies

$$0 < \underline{C} \cdot p_0 \leq p_\ell(\boldsymbol{\alpha}) \leq \overline{C} \cdot p_0 < \infty \qquad \forall \boldsymbol{\alpha} \in \mathbb{R}^{d_\ell}, \quad \|\boldsymbol{\alpha}\|_2 = 1$$

for some constants $\underline{C}, \overline{C}$, where p_0 is the density of the uniform measure on $\{\boldsymbol{u} \in \mathbb{R}^{d_\ell} : \|\boldsymbol{u}\|_2 = 1\}$.

Remark 9.1 Our non-uniform semi-random model ensures that $\|\boldsymbol{y}_i\|_2 = 1$ for all $i \in [N]$, a common normalizing assumption made in previous works on sparse subspace clustering [SC12, SEC14, WX16]. However, such a property is only used in our theoretical analysis, and in our CoCoLasso algorithm the norms of $\{y_i\}_{i=1}^N$ are assumed unknown. Indeed, if the exact norms of $\|\boldsymbol{y}_i\|_2$ are known to the data analyst the sample complexity in our analysis can be further improved, as we remarked in Remark 9.3.

The non-uniform semi-random model considers fixed (deterministic) subspaces $\{\mathcal{S}_\ell\}$, but assumes that data points within each low-dimensional subspace are independently generated from an unknown distribution P_ℓ with densities bounded away and above from below. This helps simplifying the “inter-subspace incoherence” (Definition 9.6) in our proof and yields interpretable results.

Compared with existing definitions of semi-random models [SC12, WX16, HB15, PCS14], the key difference is that in our model data are *not* uniformly distributed on each low-dimensional subspace. Instead, it is assumed that the data points are i.i.d., and that the data density is bounded away from both above and below. Such non-uniformity rules out algorithms that exploit the $\mathbb{E}[\boldsymbol{y}_i] = \mathbf{0}$ property in traditional semi-random models which is too strong and rarely holds true in practice.

Because the underlying subspaces are fixed, quantities that characterize the "affinity" between these subspace are needed because closer subspaces are harder to distinguish from each other. We adopt the following affinity measure, which was commonly used in previous works on sparse subspace clustering [WX16, WWS15a, CJW17]:

Definition 9.3 (Subspace Affinity) Let $\mathcal{U}_j$ and $\mathcal{U}_k$ be two linear subspaces of $\mathbb{R}^n$ of dimension d_j and d_k. The *affinity* between $\mathcal{U}_j$ and $\mathcal{U}_k$ is defined as $\chi_{j,k}^2 := \cos^2\theta_{jk}^{(1)} + \cdots + \cos^2\theta_{jk}^{(\min(d_j,d_k))}$, where $\theta_{jk}^{(\ell)}$ is the ℓth canonical angle between $\mathcal{U}_j$ and $\mathcal{U}_k$.

Remark 9.2 $\chi_{jk} = \|\mathbf{U}_j^\top \mathbf{U}_k\|_F$, where $\mathbf{U}_j \in \mathbb{R}^{n\times d_j}, \mathbf{U}_k \in \mathbb{R}^{n\times d_k}$ are orthonormal basis of $\mathcal{U}_j, \mathcal{U}_k$.

Throughout the paper we also write $\chi := \max_{j\neq k}\chi_{j,k}$.

For the missing data model, we need the following additional "inner-subspace" incoherence of the subspaces to ensure that the observed data entries contain sufficient amount of information. Such incoherence assumptions were widely adopted in the matrix completion community [CR09, KMO10, Rec11].

Definition 9.4 (Inner-Subspace Incoherence) Fix $\ell \in [L]$ and let $\mathbf{U}_\ell \in \mathbb{R}^{n\times d_\ell}$ be an orthonormal basis of subspace $\mathcal{U}_\ell$. The *subspace incoherence* of $\mathcal{U}_\ell$ is the smallest μ_ℓ such that

$$\max_{1\leq i\leq n} \|\boldsymbol{e}_i^\top \mathbf{U}_\ell\|_2^2 \leq \mu_\ell d_\ell / n.$$

With the above definitions, we are now ready to state the following two theorems which give sufficient success conditions for the self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^n$ produced by CoCoLasso.

Theorem 9.1 (The Gaussian Noise Model) *Suppose* $\lambda \asymp 1/\sqrt{d}$ *and* $\boldsymbol{\Delta}_{jk} \asymp \sigma^2\sqrt{\frac{\log N}{n}}$ *for all* $j, k \in [N]$. *Suppose also that* $N_\ell \geq 2\overline{C}d_\ell/\underline{C}$. *There exists a constant* $K_0 > 0$ *such that, if*

$$\sigma < K_0\left(n/d^3\log^2(\overline{C}N/\underline{C})\right)^{1/4},$$

then the optimal solution $\{\boldsymbol{c}_i\}_{i=1}^N$ *of the* CoCoSSC *estimator satisfies the subspace detection property (SDP) with probability* $1 - O(N^{-10})$.

Theorem 9.2 (The Missing Data Model) *Suppose* $\lambda \asymp 1/\sqrt{d}$, $\boldsymbol{\Delta}_{jk} \asymp \frac{\mu d\log N}{\rho\sqrt{n}}$ *for* $j \neq k$ *and* $\boldsymbol{\Delta}_{jk} \asymp \frac{\mu d\log N}{\rho^{3/2}\sqrt{n}}$ *for* $j = k$. *Suppose also that* $N_\ell \geq 2\overline{C}d_\ell/\underline{C}$. *There exists a constant* $K_1 > 0$ *such that, if*

$$\rho > K_1 \max\left\{(\mu\chi d^{5/2}\log^2 N)^{2/3}\cdot n^{-1/3}, (\mu^2 d^{7/2}\log^2 N)^{2/5}\cdot n^{-2/5}\right\},$$

then the optimal solution $\{\boldsymbol{c}_i\}_{i=1}^N$ *of the* CoCoSSC *estimator satisfies the subspace detection property (SDP) with probability* $1 - O(N^{-10})$.

Remark 9.3 If the norms of the data points $\|\boldsymbol{y}_i\|_2$ are exactly known and can be explicitly used in algorithm design, the diagonal terms of $\mathbf{A}$ in Eq. (4.1) can be directly set to $\mathbf{A}_{ii} = \|\boldsymbol{y}_i\|_2^2$ in order to avoid the ψ_2 concentration term in our proof (Definition 9.5). This would improve the sample complexity in the success condition to $\rho > \Omega(n^{-1/2})$, matching the sample complexity in linear regression problems with missing design entries [WWBS17].

Theorems 9.1 and 9.2 show that when the noise magnitude (σ in the Gaussian noise model and ρ^{-1} in the missing data model) is sufficiently small, a careful choice of tuning parameter λ results in a self-similarity matrix $\{\boldsymbol{c}_i\}$ satisfying the subspace detection property. Furthermore, the maximum amount of noise our method can tolerate is $\sigma = O(n^{1/4})$ and $\rho = \Omega(\chi^{2/3}n^{-1/3} + n^{-2/5})$, which improves over the sample complexity of existing methods (see Table 4.1).

9.1.2 The Fully Random Model

When the underlying subspaces $\mathcal{U}_1, \cdots, \mathcal{U}_L$ are independently uniformly sampled, a model referred to as the *fully random* model in the literature [SC12, SEC14, WX16], the success condition in Theorem 9.2 can be further simplified:

Corollary 9.1 *Suppose subspaces* $\mathcal{U}_1, \cdots, \mathcal{U}_L$ *have the same intrinsic dimension* d *and are uniformly sampled, the condition in Theorem 9.2 can be simplified to*

$$\rho > \widetilde{K}_1(\mu^2 d^{7/2} \log^2 N)^{2/5} \cdot n^{-2/5},$$

where $\widetilde{K}_1 > 0$ *is a new universal constant.*

Corollary 9.1 shows that in the fully random model, the $\chi^{2/3}n^{-1/3}$ term in Theorem 9.2 is negligible and the success condition becomes $\rho = \Omega(n^{-2/5})$, strictly improving existing results (see Table 4.1).

9.2 Proofs

In this section we give proofs of our main results. Due to space constraints, we only give a proof framework and leave the complete proofs of all technical lemmas to the appendix.

9.2.1 Noise Characterization and Feasibility of Pre-Processing

Definition 9.5 (Characterization of Noise Variables) $\{z_i\}$ are independent random variables and $\mathbb{E}[z_i] = \mathbf{0}$. Furthermore, there exist parameters $\psi_1, \psi_2 > 0$ such that with probability $1 - O(N^{-10})$ the following holds uniformly for all $i, j \in [N]$:

$$\left|z_i^\top \boldsymbol{y}_j\right| \le \psi_1 \sqrt{\frac{\log N}{n}}; \qquad \left|z_i^\top z_j - \mathbb{E}[z_i^\top z_j]\right| \le \begin{cases} \psi_1 \sqrt{\frac{\log N}{n}} & i \ne j; \\ \psi_2 \sqrt{\frac{\log N}{n}} & i = j. \end{cases}$$

Proposition 9.1 *Suppose $\boldsymbol{\Delta}$ are set as $\boldsymbol{\Delta}_{jk} \ge 3\psi_1 \sqrt{\frac{\log N}{n}}$ for $j \ne k$ and $\boldsymbol{\Delta}_{jk} \ge 3\psi_2 \sqrt{\frac{\log N}{n}}$ for $j = k$. Then with probability $1 - O(N^{-10})$ the set S defined in Eq. (4.1) is not empty.*

The following two lemmas derive explicit bounds on ψ_1 and ψ_2 for the two noise models.

Lemma 9.1 *The Gaussian noise model satisfies Definition 9.5 with $\psi_1 \lesssim \sigma^2$ and $\psi_2 \lesssim \sigma^2$.*

Lemma 9.2 *Suppose $\rho = \Omega(n^{-1/2})$. The missing data model satisfies Definition 9.5 with $\psi_1 \lesssim \rho^{-1}\mu d\sqrt{\log N}$ and $\psi_2 \lesssim \rho^{-3/2}\mu d\sqrt{\log N}$, where $d = \max_{\ell \in [L]} d_\ell$ and $\mu = \max_{\ell \in [L]} \mu_\ell$.*

9.2.2 Optimality Condition and Dual Certificates

We first write down the dual problem of CoCoSSC:

$$\text{Dual CoCoSSC}: \qquad \boldsymbol{\nu}_i = \arg\max_{\boldsymbol{\nu}_i \in \mathbb{R}^N} \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\nu}_i - \frac{1}{2\lambda}\|\boldsymbol{\nu}_i\|_2^2 \quad s.t. \quad \left\|\widetilde{\mathbf{X}}_{-i}^\top \boldsymbol{\nu}_i\right\|_\infty \le 1. \tag{9.1}$$

Lemma 9.3 (Dual Certificate, Lemma 12 of [WX16]) *Suppose there exists triplet $(\boldsymbol{c}, \boldsymbol{e}, \boldsymbol{\nu})$ such that $\tilde{\boldsymbol{x}}_i = \widetilde{\mathbf{X}}_{-i}\boldsymbol{c} + \boldsymbol{e}$, $\boldsymbol{c}$ has support $S \subseteq T \subseteq [N]$, and that $\boldsymbol{\nu}$ satisfies*

$$[\widetilde{\mathbf{X}}_{-i}]_S^\top \boldsymbol{\nu} = \operatorname{sgn}(\boldsymbol{c}_S), \quad \boldsymbol{\nu} = \lambda \boldsymbol{e}, \quad \left\|[\widetilde{\mathbf{X}}_{-i}]_{T \cap S^c}^\top \boldsymbol{\nu}\right\|_\infty \le 1, \quad \left\|[\widetilde{\mathbf{X}}_{-i}]_{T^c}^\top \boldsymbol{\nu}\right\|_\infty < 1,$$

then any optimal solution $\boldsymbol{c}_i$ to Eq. (4.2) satisfies $[\boldsymbol{c}_i]_{T^c} = \mathbf{0}$.

To construct such a dual certificate and to de-couple potential statistical dependency, we follow [WX16] to consider a constrained version of the optimization problem. Let $\widetilde{\mathbf{X}}_{-i}^{(\ell)}$ denote the data matrix of all but $\tilde{\boldsymbol{x}}_i$ in cluster $\mathcal{S}_\ell$. The constrained problems are defined as follows:

$$\text{Constrained Primal :} \quad \tilde{\boldsymbol{c}}_i = \arg\min_{\boldsymbol{c}_i \in \mathbb{R}^{N_\ell - 1}} \|\boldsymbol{c}_i\|_1 + \lambda/2 \cdot \|\tilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)} \boldsymbol{c}_i\|_2^2; \tag{9.2}$$

$$\text{Constrained Dual :} \quad \tilde{\boldsymbol{\nu}}_i = \arg\max_{\boldsymbol{\nu}_i \in \mathbb{R}^{N_\ell - 1}} \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\nu}_i - 1/(2\lambda) \cdot \|\boldsymbol{\nu}_i\|_2^2 \quad s.t. \quad \|(\widetilde{\mathbf{X}}_{-i}^{(\ell)})^\top \boldsymbol{\nu}_i\|_\infty \le 1. \tag{9.3}$$

With $\boldsymbol{c} = [\tilde{\boldsymbol{c}}_i, \mathbf{0}_{\mathcal{S}_{-\ell}}]$, $\boldsymbol{\nu} = [\tilde{\boldsymbol{\nu}}_i, \mathbf{0}_{\mathcal{S}_{-\ell}}]$, and $\boldsymbol{e} = \tilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)} \tilde{\boldsymbol{c}}_i$, the certificate satisfies the first three conditions in Lemma 9.3 with $T = \mathcal{S}_\ell$ and $S = \text{supp}(\tilde{\boldsymbol{c}}_i)$. Therefore, we only need to establish that $|\langle \tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{\nu}}_i \rangle| < 1$ for all $\tilde{\boldsymbol{x}}_j \notin \mathcal{S}_\ell$ to show no false discoveries, which we prove in the next section.

9.2.3 Deterministic Success Conditions

Define the following deterministic quantities as *inter-subspace incoherence* and *in-radius*, which are important quantities in deterministic analysis of sparse subspace clustering methods [SC12, WX16, SEC14].

Definition 9.6 (Inter-Subspace Incoherence) The inter-subspace incoherence $\tilde{\mu}$ is defined as $\tilde{\mu} := \max_{\ell \in [L]} \max_{\boldsymbol{y}_i \in \mathcal{S}_\ell} \max_{\boldsymbol{y}_j \notin \mathcal{S}_\ell} \left|\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle\right|$.

Definition 9.7 (In-Radius) Define r_i as the radius of the largest ball inscribed in the convex body of $\{\pm \mathbf{Y}_{-j}^{(\ell)}\}$. Also define that $r := \min_{1 \le i \le N} r_i$.

The following lemma derives an upper bound on $|\langle \tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{\nu}}_i \rangle|$, which is proved in the appendix.

Lemma 9.4 *For every (i, j) belonging to different clusters, $|\langle \tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{\nu}}_i \rangle| \lesssim \lambda(1 + \|\tilde{\boldsymbol{c}}_i\|_1)(\tilde{\mu} + \psi_1 \sqrt{\log N/n})$, where $\|\tilde{\boldsymbol{c}}_i\|_1 \lesssim r^{-1}(1 + r^{-1}\lambda(\psi_1 + \psi_2)\sqrt{\log N/n})$.*

Lemmas 9.3 and 9.4 immediately yield the following theorem:

Theorem 9.3 (No False Discoveries) *There exists an absolute constant $\kappa_1 > 0$ such that if*

$$\frac{\lambda}{r}\left(1 + \frac{\lambda}{r}(\psi_1 + \psi_2)\sqrt{\frac{\log N}{n}}\right) \cdot \left(\tilde{\mu} + \psi_1 \sqrt{\frac{\log N}{n}}\right) < \kappa_1, \tag{9.4}$$

then the optimal solution $\boldsymbol{c}_i$ of the CoCoSSC *estimator in Eq. (4.2) has no false discoveries, that is, $\boldsymbol{c}_{ij} = 0$ for all $\boldsymbol{x}_j$ that belongs to a different cluster of $\boldsymbol{x}_i$.*

The following theorem shows conditions under which $\boldsymbol{c}_i$ is not the trivial solution $\boldsymbol{c}_i = \mathbf{0}$.

Theorem 9.4 (Avoiding Trivial Solutions) *There exists an absolute constant* $\kappa_2 > 0$ *such that, if*

$$\lambda\left(r - \psi_1\sqrt{\frac{\log N}{n}}\right) > \kappa_2, \tag{9.5}$$

then the optimal solution $\mathbf{c}_i$ *of the* CoCoSSC *estimator in Eq. (4.2) is non-trivial, that is,* $\mathbf{c}_i \neq \mathbf{0}$.

Finally, we remark that choosing $r = c/\lambda$ for some small constant $c > 0$ (depending only on κ_1 and κ_2), the choice of λ satisfies both theorems 9.3 and 9.4 provided that

$$\max\left\{\frac{\psi_1}{r}\sqrt{\frac{\log N}{n}}, \frac{\tilde{\mu}}{r^2}, \frac{\tilde{\mu}(\psi_1+\psi_2)}{r^3}\sqrt{\frac{\log N}{n}}, \frac{\psi_1(\psi_1+\psi_2)}{r^3}\frac{\log N}{n}\right\} < \kappa_3 \tag{9.6}$$

for some sufficiently small absolute constant $\kappa_3 > 0$ that depends on κ_1, κ_2, and c.

9.2.4 Bounding $\tilde{\mu}$ and r in Randomized Models

Lemma 9.5 *Suppose* $N_\ell = \Omega(\overline{C}d_\ell/\underline{C}_\ell)$. *Under the non-uniform semi-random model, with probability* $1 - O(N^{-10})$ *it holds that* $\tilde{\mu} \lesssim \chi\sqrt{\log(\overline{C}N/\underline{C})}$ *and* $r \gtrsim 1/\sqrt{d}$.

Lemma 9.6 *Suppose* $\mathcal{U}_1, \ldots, \mathcal{U}_L$ *are independently uniformly sampled linear subspaces of dimension* d *in* $\mathbb{R}^n$. *Then with probability* $1 - O(N^{-10})$ *we have that* $\chi \lesssim d\sqrt{\log N/n}$ *and* $\mu \lesssim \sqrt{\log N}$.

9.3 Numerical Results

Experimental Settings and Methods We conduct numerical experiments based on synthetic generated data, using a computer with Intel Core i7 CPU (4 GHz) and 16 GB memory. Each synthetic data set has ambient dimension $n = 100$, intrinsic dimension $d = 4$, number of underlying subspaces $L = 10$, and a total number of $N = 1000$ unlabeled data points. The observation rate ρ and Gaussian noise magnitude σ vary in our simulations. Underlying subspaces are generated uniformly at random, corresponding to our fully random model. Each data point has an equal probability of being assigned to any cluster and is generated uniformly at random on its corresponding low-dimensional subspace.

We compare the performance (explained later) of our CoCoSSC approach, and two popular existing methods Lasso SSC and the de-biased Dantzig selector. The ℓ_1 regularized self-regression steps in both CoCoSSC and Lasso SSC are implemented using ADMM. The pre-processing step of CoCoSSC is implemented using alternating projections initialized at $\tilde{\mathbf{\Sigma}} = \mathbf{X}^\top\mathbf{X} - \mathbf{D}$. Unlike the theoretical recommendations, we choose $\mathbf{\Delta}$ in Eq. (4.1) to be very large (3×10^3 for diagonal entries and 10^3 for off-diagonal entries) for fast convergence. The de-biased Dantzig selector is implemented using linear programming.

Evaluation Measure We consider two measures to evaluate the performance of algorithms being compared. The first one evaluates the quality of the similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ by measuring how far (relatively) it deviates from having the subspace detection property. In particular, we consider the RelViolation metric propositioned in [WX16] defined as

$$\text{RelViolation}(C, \mathcal{M}) = (\textstyle\sum_{(i,j)\notin\mathcal{M}}|C|_{i,j})/(\textstyle\sum_{(i,j)\in\mathcal{M}}|C|_{i,j}), \tag{9.7}$$

where $\mathcal{M}$ is the mask of ground truth with all (i, j) satisfying $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}^{(\ell)}$ for some ℓ. A high RelViolation indicates frequent deviation from the subspace detection propositionerty and therefore poorer quality of $\{\boldsymbol{c}_i\}_{i=1}^N$.

For clustering results, we use the Fowlkes–Mallows index [FM83] to evaluate their quality. Suppose $\mathcal{A} \subseteq \{(i, j) \in [N] \times [N]\}$ consists of pairs of data points that are clustered together by a clustering algorithm, and $\mathcal{A}_0$ is the ground truth clustering. Define $TP = |\mathcal{A} \cap \mathcal{A}_0|$, $FP = |\mathcal{A} \cap \mathcal{A}_0^c|$, $FN = |\mathcal{A}^c \cap \mathcal{A}_0|$, $TN = |\mathcal{A}^c \cap \mathcal{A}_0^c|$. The Fowlkes–Mallows (FM) index is then expressed as

$$FM = \sqrt{TP^2/(TP + FP)(TP + FN)}.$$

The FM index of any two clusterings $\mathcal{A}$ and $\mathcal{A}_0$ is always between 0 and 1, with an FM index of one indicating perfectly identical clusterings and an FM index close to zero otherwise.

Results We first give a qualitative illustration of similarity matrices $\{\boldsymbol{c}_i\}_{i=1}^N$ produced by the three algorithms of Lasso SSC, de-biased Dantzig selector, and CoCoSSC in Fig. 9.1. We observe that the similarity matrix of Lasso SSC has several spurious connections, and both Lasso SSC and the de-biased Dantzig selector suffer from graph connectivity issues as signals within each block (cluster) are not very strong. On the other hand, the similarity matrix of CoCoSSC produces convincing signals within each block (cluster). This shows that our propositioned CoCoSSC approach not only has few false discoveries as predicted by our theoretical results, but also has much better graph connectivity which our theory did not attempt to cover.

In Fig. 9.2 we report the Fowlkes–Mallows (FM) index for clustering results and RelViolation scores of similarity matrices $\{\boldsymbol{c}_i\}_{i=1}^N$ under various noise magnitude (σ) and observation rates (ρ) settings. A grid of tuning parameter values λ are attempted and the one leading to the best performance is reported. It is observed that our propositioned CoCoLasso consistently outperforms its competitors Lasso SSC

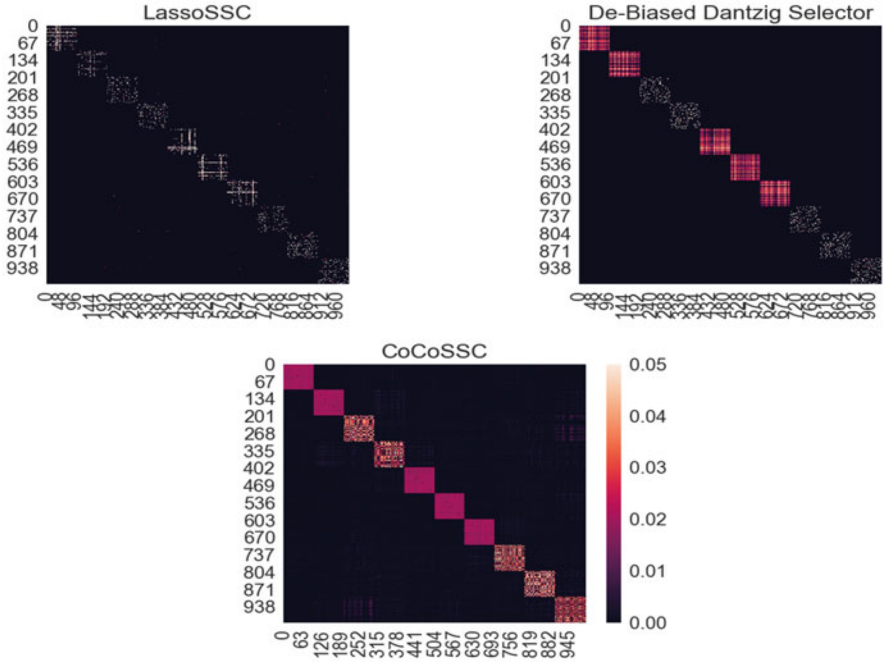

Fig. 9.1 Heatmaps of similarity matrices $\{\boldsymbol{c}_i\}_{i=1}^N$, with brighter colors indicating larger absolute values of matrix entries. Left: LassoSSC; Middle: De-biased Dantzig selector; Right: CoCoSSC

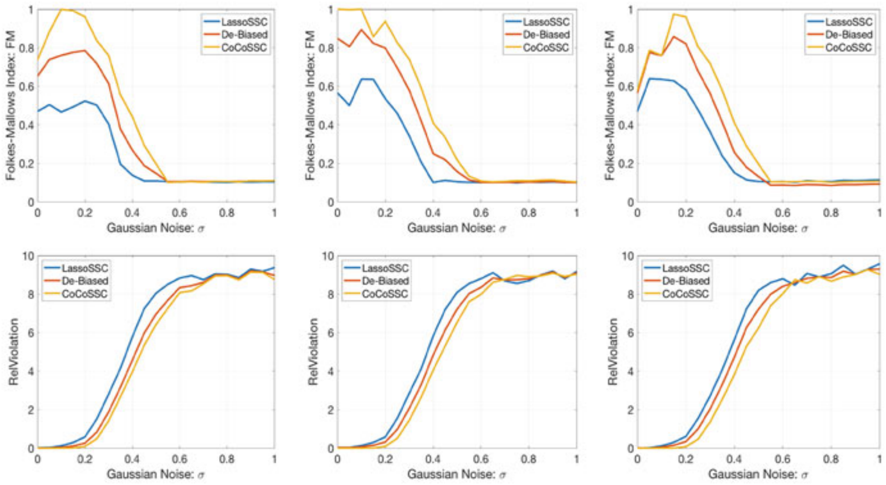

Fig. 9.2 The Fowlkes–Mallows (FM) index of clustering results (top row) and RelViolation scores (bottom row) of the three methods, with noise of magnitude σ varying from 0 to 1. Left column: missing rate $1 - \rho = 0.03$, middle column: $1 - \rho = 0.25$, right column: $1 - \rho = 0.9$

and de-biased Dantzig selector. Furthermore, CoCoLasso is very computationally efficient and converges in 8–15 seconds on each synthetic data set. On the other hand, de-biased Dantzig selector is computationally very expensive and typically takes over 100 seconds to converge.

9.4 Technical Details

Proof of Proposition 9.1 By Definition 9.5 we know that $|\widetilde{\mathbf{\Sigma}}_{-i} - \mathbf{Y}_{-i}^T\mathbf{Y}_{-i}| \le |\mathbf{\Delta}|$ in an element-wise sense. Also note that $\mathbf{Y}^\top\mathbf{Y}$ is positive semidefinite. Thus, $\mathbf{Y}^\top\mathbf{Y} \in S$. □

Proofs of Lemmas 9.1 and 9.2 Lemma 9.1 is proved in [WX16]. See Lemmas 17 and 18 of [WX16] and note that $\mathbb{E}[\boldsymbol{z}_i^\top \boldsymbol{z}_i] = \sigma^2$.

We next prove Lemma 9.2. We first consider $|\boldsymbol{z}_i^\top \boldsymbol{y}_j|$. Let $\boldsymbol{z} = \boldsymbol{z}_i$, $\boldsymbol{y} = \boldsymbol{y}_i$, $\tilde{\boldsymbol{y}} = \boldsymbol{y}_j$, and $\boldsymbol{r} = R_{j\cdot}$. Define $T_i := z_i \boldsymbol{y}_i = (1 - \boldsymbol{r}_i/\rho)\boldsymbol{y}_i\tilde{\boldsymbol{y}}_j$. Because $\boldsymbol{r}$ is independent of $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$, we have that $\mathbb{E}[T_i] = 0$, $\mathbb{E}[T_i^2] \le \boldsymbol{y}_i^2\tilde{\boldsymbol{y}}_i{}^2/\rho \le \mu^2 d^2/\rho n^2$, and $|T_i| \le \mu d/\rho n =: M$ almost surely. Using Bernstein's inequality, we know that with probability $1 - O(N^{-10})$

$$|\boldsymbol{z}_i^\top \boldsymbol{y}_j| = \left|\sum_{i=1}^{T} T_i\right| \lesssim \sqrt{\sum_{i=1}^{n} \mathbb{E}[T_i^2] \cdot \log N} + M\log N \lesssim \mu d\sqrt{\frac{\log^2 N}{\rho n}}.$$

We next consider $|\boldsymbol{z}_i^\top \boldsymbol{z}_j|$ and the $i \neq j$ case. Let $\boldsymbol{y} = \boldsymbol{y}_i$, $\tilde{\boldsymbol{y}} = \boldsymbol{y}_j$, $\boldsymbol{r} = R_{i\cdot}$, and $\tilde{\boldsymbol{r}} = R_{j\cdot}$. By definition of μ, we have that $\|\boldsymbol{y}\|_\infty^2 \le \mu d_i/n$ and $\|\tilde{\boldsymbol{y}}\|_\infty^2 \le \mu d_j/n$. Define $T_i := z_i\tilde{z}_i = (1 - \boldsymbol{r}_i/\rho)(1 - \tilde{\boldsymbol{r}}_i/\rho) \cdot \boldsymbol{y}_i\tilde{\boldsymbol{y}}_i$. Because $\boldsymbol{r}$ and $\tilde{\boldsymbol{r}}$ are independent, $\mathbb{E}[T_i] = 0$, $\mathbb{E}[T_i^2] \le \boldsymbol{y}_i^2\tilde{\boldsymbol{y}}_i^2/\rho^2 \le \mu^2 d^2/\rho^2 n^2$, and $|T_i| \le \mu d/\rho^2 n =: M$ almost surely. Using Bernstein's inequality, we know that with probability $1 - O(N^{-10})$

$$\left|\sum_{i=1}^{n} T_i\right| \lesssim \sqrt{\sum_{i=1}^{n} \mathbb{E}[T_i^2] \cdot \log N} + M\log N \lesssim \frac{\mu d}{\rho}\sqrt{\frac{\log^2 N}{n}},$$

where the last inequality holds because $\rho = O(n^{-1/2})$.

Finally is the case of $|\boldsymbol{z}_i^\top \boldsymbol{z}_j|$ and $i = j$. Let again $\boldsymbol{z} := \boldsymbol{z}_i = \boldsymbol{z}_j$. Define $T_i := \boldsymbol{z}_i^2 - \mathbb{E}[\boldsymbol{z}_i^2] = (1 - \boldsymbol{r}_i/\rho)^2\boldsymbol{y}_i^2 - (1-\rho)^2/\rho \cdot \boldsymbol{y}_i^2$. It is easy to verify that $\mathbb{E}[T_i] = 0$, $\mathbb{E}[T_i^2] \lesssim \boldsymbol{y}_i^4/\rho^3 \le \mu^2 d^2/\rho^3 n^2$, and $|T_i| \lesssim \boldsymbol{y}_i^2/\rho^2 \le \mu d/\rho^2 n$. Subsequently, with probability $1 - O(N^{-10})$ we have

$$\left|\sum_{i=1}^{n} T_i\right| \lesssim \frac{\mu d}{\rho^{3/2}}\sqrt{\frac{\log^2 N}{n}}.$$

The estimation error of $(1-\rho)(\mathbf{X}^\top\mathbf{X})_{ii}$ for $(1-\rho)/\rho \cdot \|\boldsymbol{y}_i\|_2^2 = (1-\rho)/\rho$ can be upper bounded similarly. □

Proof of Lemma 9.4 Take $\boldsymbol{\Delta}_{jk} = 3\psi_1\sqrt{\frac{\log N}{n}}$ for $j \neq k$ and $\boldsymbol{\Delta}_{jk} = 3\psi_2\sqrt{\frac{\log N}{n}}$. Fix arbitrary $\tilde{\boldsymbol{x}}_j \notin \mathcal{S}_\ell$ and $\tilde{\boldsymbol{x}}_i \in \mathcal{S}_\ell$. Because $\tilde{\boldsymbol{\nu}}_i = \lambda(\tilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}^{(\ell)}_{-i}\tilde{\boldsymbol{c}}_i)$, we have that

$$\begin{aligned}
\left|\langle \tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{\nu}}_i\rangle\right| = \lambda\left|\tilde{\boldsymbol{x}}_j^\top(\tilde{\boldsymbol{x}}_i + \widetilde{\mathbf{X}}^{(\ell)}_{-i}\tilde{\boldsymbol{c}}_i)\right| &\le \lambda(1+\|\tilde{\boldsymbol{c}}_i\|_1)\cdot \sup_{\tilde{\boldsymbol{x}}_i \in \mathcal{S}_\ell}\left|\langle \tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_i\rangle\right| \\
&\le \lambda(1+\|\tilde{\boldsymbol{c}}_i\|_1)\cdot\left(\tilde{\mu} + \sup_{\tilde{\boldsymbol{x}}_i \notin \mathcal{S}_\ell}\left|\langle \tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_i\rangle - \langle \boldsymbol{y}_j, \boldsymbol{y}_i\rangle\right|\right) \\
&\lesssim \lambda(1+\|\tilde{\boldsymbol{c}}_i\|_1)\cdot\left(\tilde{\mu} + \psi_1\sqrt{\frac{\log N}{n}}\right),
\end{aligned} \tag{9.8}$$

where the last inequality holds by applying Definition 9.5 and the fact that

$$\begin{aligned}
\left|\langle \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle - \langle \tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_j\rangle\right| &\le \left|(\widetilde{\boldsymbol{\Sigma}}_+)_{ij} - (\widetilde{\boldsymbol{\Sigma}})_{ij}\right| + \left|(\widetilde{\boldsymbol{\Sigma}})_{ij} - \langle \tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_j\rangle\right| \\
&\le \left|\boldsymbol{\Delta}_{ij}\right| + \left|\langle \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle - \langle \tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_j\rangle\right| \\
&\le \left|\boldsymbol{\Delta}_{ij}\right| + \left|\langle \tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{y}}_j\rangle\right| + \left|\langle \tilde{\boldsymbol{y}}_j, \tilde{\boldsymbol{z}}_i\rangle\right| + \left|\langle \tilde{\boldsymbol{z}}_j, \tilde{\boldsymbol{z}}_i\rangle\right| \\
&\lesssim \psi_1\sqrt{\frac{\log N}{n}} \quad \text{for } i \neq j.
\end{aligned}$$

To bound $\|\tilde{\boldsymbol{c}}_i\|_1$, consider an auxiliary noiseless problem:

$$\hat{\boldsymbol{c}}_i := \arg\min_{\boldsymbol{c}_i}\|\boldsymbol{c}_i\|_1 \quad s.t. \;\; \boldsymbol{y}_i = \mathbf{Y}^{(\ell)}_{-i}\boldsymbol{c}_i. \tag{9.9}$$

Note that when $r > 0$ Eq. (9.9) is always feasible. Following standard analysis (e.g., Lemma 15 and Eq. (5.15) of [WX16]), it can be established that $\|\hat{\boldsymbol{c}}_i\|_1 \le 1/r_i \le 1/r$. On the other hand, by optimality we have $\|\tilde{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\|\tilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}^{(\ell)}_{-i}\tilde{\boldsymbol{c}}_i\|_2^2 \le \|\hat{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\|\tilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}^{(\ell)}_{-i}\hat{\boldsymbol{c}}_i\|_2^2$. Therefore,

$$\begin{aligned}
\|\tilde{\boldsymbol{c}}_i\|_1 &\le \|\hat{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\left\|\tilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}^{(\ell)}_{-i}\hat{\boldsymbol{c}}_i\right\|_2^2 \\
&\lesssim \|\hat{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\left\|\boldsymbol{y}_i - \mathbf{Y}^{(\ell)}_{-i}\hat{\boldsymbol{c}}_i\right\|_2^2 + (1+\|\hat{\boldsymbol{c}}_i\|_1)^2\cdot\frac{\lambda}{2}\sup_{\boldsymbol{y}_i,\boldsymbol{y}_j \in \mathcal{S}_\ell}\left|\langle \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle - \langle \boldsymbol{y}_i, \boldsymbol{y}_j\rangle\right| \\
&= \|\hat{\boldsymbol{c}}_i\|_1 + (1+\|\hat{\boldsymbol{c}}_i\|_1)^2\cdot\frac{\lambda}{2}\sup_{\boldsymbol{y}_i,\boldsymbol{y}_j \in \mathcal{S}_\ell}\left|\langle \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle - \langle \boldsymbol{y}_i, \boldsymbol{y}_j\rangle\right|
\end{aligned}$$

$$\begin{aligned}&\lesssim \|\hat{\boldsymbol{c}}_i\|_1 + (1+\|\hat{\boldsymbol{c}}_i\|_1)^2 \cdot (\psi_1+\psi_2)\sqrt{\frac{\log N}{n}}\\ &\lesssim \frac{1}{r}\left(1+\frac{\lambda}{r}(\psi_1+\psi_2)\sqrt{\frac{\log N}{n}}\right). \end{aligned}\tag{9.10}$$

□

Proof of Theorem 9.4 Following the analysis of Lasso SSC solution path in [WX16], it suffices to show that $\lambda > 1/\|\tilde{\boldsymbol{x}}_i^\top \widetilde{\mathbf{X}}_{-i}\|_\infty$. On the other hand, note that $\|\boldsymbol{y}_i^\top \mathbf{Y}_{-i}\|_\infty \geq \|\boldsymbol{y}_i^\top \mathbf{Y}_{-i}^{(\ell)}\|_\infty \geq r_i \geq r$ (see, for example, Eq. (5.19) of [WX16]). Subsequently,

$$\left\|\tilde{\boldsymbol{x}}_i^\top \widetilde{\mathbf{X}}_{-i}\right\|_\infty \geq \left\|\boldsymbol{y}_i^\top \mathbf{Y}_{-i}\right\|_\infty - \sup_{j\neq i}\left|\langle \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle - \langle \boldsymbol{y}_i, \boldsymbol{y}_j\rangle\right| \gtrsim r - \psi_1\sqrt{\frac{\log N}{n}}.$$

□

Proof of Lemma 9.5 We first prove

$$\max_{\boldsymbol{y}_i\in\mathcal{S}_k}\max_{\boldsymbol{y}_j\in\mathcal{S}_\ell}\left|\langle \boldsymbol{y}_i, \boldsymbol{y}_j\rangle\right| \lesssim \chi_{k\ell}\cdot\frac{\log(\overline{C}N/\underline{C})}{\sqrt{d_k d_\ell}} \qquad \forall j\neq k\in[L]. \tag{9.11}$$

Let N_k and N_ℓ be the total number of data points in $\mathcal{S}_k$ and $\mathcal{S}_\ell$, and let P_k and P_ℓ be the corresponding densities which are bounded from both above and below by $\overline{C}p_0$ and $\underline{C}p_0$. Consider a rejection sampling procedure: first sample $\boldsymbol{\alpha}$ randomly from the uniform measure over $\{\boldsymbol{\alpha}\in\mathbb{R}^{d_k} : \|\boldsymbol{\alpha}\|_2 = 1\}$, and then reject the sample if $u > p_k(\boldsymbol{\alpha})/\overline{C}p_0$, where $u\sim U(0,1)$. Repeat the procedure until N_k samples are obtained. This procedure is sound because $p_k/p_0 \leq \overline{C}$, and the resulting (accepted) samples are i.i.d. distributed according to P_k. On the other hand, for any $\boldsymbol{\alpha}$ the probability of acceptance is lower bounded by $\underline{C}/\overline{C}$. Therefore, the procedure terminates by producing a total of $O(\overline{C}N_k/\underline{C})$ samples (both accepted and rejected). Thus, without loss of generality we can assume both P_k and P_ℓ are uniform measures on the corresponding spheres, by paying the cost of adding $\tilde{N}_k = O(\overline{C}N_k/\underline{C})$ and $\tilde{N}_\ell = O(\overline{C}N_\ell/\underline{C})$ points to each subspace.

Now fix $\boldsymbol{y}_i = \mathbf{U}_k\boldsymbol{\alpha}_i$ and $\boldsymbol{y}_j = \mathbf{U}_\ell\boldsymbol{\alpha}_j$, where $\boldsymbol{\alpha}_i\in\mathbb{R}^{d_k}$, $\boldsymbol{\alpha}_j\in\mathbb{R}^{d_\ell}$, and $\|\boldsymbol{\alpha}_i\|_2 = \|\boldsymbol{\alpha}_j\|_2 = 1$. Then both $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ are uniformly distributed on the low-dimensional spheres, and that $|\langle \boldsymbol{y}_i, \boldsymbol{y}_j\rangle| = |\boldsymbol{\alpha}_i^\top(\mathbf{U}_k^\top\mathbf{U}_\ell)\boldsymbol{\alpha}_j|$. Applying Lemma 7.5 of [SC12] and note that $\chi_{k\ell} = \|\mathbf{U}_k^\top\mathbf{U}_\ell\|_F$ we complete the proof of Eq. (9.11).

We next prove

$$r_i \gtrsim \sqrt{\frac{\log(\underline{C}N_\ell/\overline{C}d_\ell)}{d_\ell}} \qquad \forall i\in[N], \ell\in[L], \boldsymbol{x}_i\in\mathcal{S}_\ell. \tag{9.12}$$

Let P_ℓ be the underlying measure of subspace $\mathcal{S}_\ell$. Consider the decomposition $P_\ell = \underline{C}/\overline{C} \cdot P_0 + (1 - \underline{C}/\overline{C}) \cdot P'_\ell$, where P_0 is the uniform measure. Such a decomposition and the corresponding density P'_ℓ exist because $\underline{C}P_0 \leq P_\ell \leq \overline{C}P_0$. This shows that the distribution of points in subspace $\mathcal{S}_\ell$ can be expressed as a mixture distribution, with a uniform density mixture with weight probability $\underline{C}/\overline{C}$. Because r_i decreases with smaller data set, it suffices to consider only the uniform mixture. Thus, we can assume P_ℓ is the uniform measure at the cost of considering only $\tilde{N}_\ell = \Omega(\underline{C}N_\ell/\overline{C})$ points in subspace $\mathcal{S}_\ell$. Applying Lemma 21 of [WX16] and replacing N_ℓ with $\tilde{N}_\ell$ we complete the proof of Eq. (9.12).

Finally Lemma 9.5 is an easy corollary of Eqs. (9.11) and (9.12). □

Proof of Lemma 9.6 Fix $k, \ell \in [L]$ and let $\mathbf{U}_k = (\boldsymbol{u}_{k1}, \cdots, \boldsymbol{u}_{kd})$, $\mathbf{U}_\ell = (\boldsymbol{u}_{\ell 1}, \cdots, \boldsymbol{u}_{\ell d})$ be orthonormal basis of $\mathcal{U}_k$ and $\mathcal{U}_\ell$. Then $\chi_{k\ell} = \|\mathbf{U}_k^\top \mathbf{U}_\ell\|_F \leq d\|\mathbf{U}_k^\top \mathbf{U}_\ell\|_{\max} = d \cdot \sup_{1\leq i,j\leq d} |\langle \boldsymbol{u}_{ki}, \boldsymbol{u}_{\ell j}\rangle|$. Because $\mathcal{U}_k$ and $\mathcal{U}_\ell$ are random subspaces, $\boldsymbol{u}_{ki}$ and $\boldsymbol{u}_{\ell j}$ are independent vectors distributed uniformly on the d-dimensional unit sphere. Applying Lemma 17 of [WX16] and a union bound over all i, j, k, ℓ we prove the upper bound on χ. For the upper bound on μ, simply note that $\|\boldsymbol{u}_{jk}\|_\infty \lesssim \sqrt{\frac{\log N}{n}}$ with probability $1 - O(N^{-10})$ by standard concentration result for Gaussian suprema. □

9.5 Concluding Remarks

Our numerical simulations first demonstrated the spectral clustering accuracy with respect to the effect of Gaussian noise. In this experiment, ambient dimension $n = 100$, intrinsic dimension $d = 4$, the number of clusters $L = 10$, the number of data points $N = 1000$, and the Gaussian noise is $Z_i j \ N(0, \sigma/\sqrt{n}$, where σ is changed from 0.00 to 1.00 with step length 0.01.

The second experiments investigated the RelViolation with respect to Gaussian noise σ and missing rate ρ. We change σ from 0 to 1 with step length 0.01 and set ρ as 0.03, 0.05, and 0.10, respectively. In these experiments, ambient dimension $n = 10$, intrinsic dimension $d = 2$, the number of clusters $L = 5$, and the number of data points $N = 100$.

Our last numerical simulations test the effects of Gaussian noise σ, subspace rand d, and number of clusters L, respectively.

An interesting future direction is to further improve the sample complexity to $\rho = \Omega(n^{-1/2})$ without knowing the norms $\|\boldsymbol{y}_i\|_2$. Such sample complexity is likely to be optimal because it is the smallest observation rate under which off-diagonal elements of sample covariance $\mathbf{X}^\top\mathbf{X}$ can be consistently estimated in max norm, which is also shown to be optimal for related regression problems [WWBS17].