Minimax Localization of Structural Information in Large Noisy Matrices

Mladen Kolar[†*]
mladenk@cs.cmu.edu

Sivaraman Balakrishnan[†*]
sbalakri@cs.cmu.edu

Alessandro Rinaldo[††]
arinaldo@stat.cmu.edu

Aarti Singh[†]
aarti@cs.cmu.edu

[†] **School of Computer Science and** [††] **Department of Statistics, Carnegie Mellon University**

Abstract

We consider the problem of identifying a sparse set of relevant columns and rows in a large data matrix with highly corrupted entries. This problem of identifying groups from a collection of bipartite variables such as proteins and drugs, biological species and gene sequences, malware and signatures, etc is commonly referred to as biclustering or co-clustering. Despite its great practical relevance, and although several ad-hoc methods are available for biclustering, theoretical analysis of the problem is largely non-existent. The problem we consider is also closely related to structured multiple hypothesis testing, an area of statistics that has recently witnessed a flurry of activity. We make the following contributions

1. We prove lower bounds on the minimum signal strength needed for successful recovery of a bicluster as a function of the noise variance, size of the matrix and bicluster of interest.
2. We show that a combinatorial procedure based on the scan statistic achieves this optimal limit.
3. We characterize the SNR required by several computationally tractable procedures for biclustering including element-wise thresholding, column/row average thresholding and a convex relaxation approach to sparse singular vector decomposition.

1 Introduction

Biclustering is the problem of identifying a (typically) sparse set of relevant columns and rows in a large, noisy data matrix. This problem along with the first algorithm to solve it were proposed by Hartigan [14] as a way to directly cluster data matrices to produce clusters with greater interpretability. Biclustering routinely arises in several applications such as discovering groups of proteins and drugs that interact with each other [19], learning phylogenetic relationships between different species based on alignments of snippets of their gene sequences [30], identifying malware that have similar signatures [7] and identifying groups of users with similar tastes for commercial products [29]. In these applications, the data matrix is often indexed by (object, feature) pairs and the goal is to identify clusters in this set of bipartite variables.

In standard clustering problems, the goal is only to identify meaningful groups of objects and the methods typically use the entire feature vector to define a notion of similarity between the objects.

---

[*]These authors contributed equally to this work

Biclustering can be thought of as high-dimensional clustering where only a subset of the features are relevant for identifying similar objects, and the goal is to identify not only groups of objects that are similar, but also which features are relevant to the clustering task. Consider, for instance gene expression data where the objects correspond to genes, and the features correspond to their expression levels under a variety of experimental conditions. Our present understanding of biological systems leads us to expect that subsets of genes will be co-expressed only under a small number of experimental conditions. Although, pairs of genes are not expected to be similar under *all* experimental conditions it is critical to be able to discover local expression patterns, which can for instance correspond to joint participation in a particular biological pathway or process. Thus, while clustering aims to identify *global* structure in the data, biclustering take a more *local* approach by jointly clustering *both* objects and features.

Prevalent techniques for finding biclusters are typically heuristic procedures with little or no theoretical underpinning. In order to study, understand and compare biclustering algorithms we consider a simple theoretical model of biclustering [18, 17, 26]. This model is akin to the spiked covariance model of [15] widely used in the study of PCA in high-dimensions.

We will focus on the following simple observation model for the matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$:

$$\mathbf{A} = \beta \mathbf{u} \mathbf{v}' + \boldsymbol{\Delta} \tag{1}$$

where $\boldsymbol{\Delta} = \{\Delta_{ij}\}_{i \in [n_1], j \in [n_2]}$ is a random matrix whose entries are i.i.d. $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$ known, $\mathbf{u} = \{u_i : i \in [n_1]\}$ and $\mathbf{v} = \{v_i : i \in [n_2]\}$ are unknown deterministic unit vectors in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively, and $\beta > 0$ is a constant. To simplify the presentation, we assume that $\mathbf{u} \propto \{0, 1\}^{n_1}$ and $\mathbf{v} \propto \{0, 1\}^{n_2}$. Let $K_1 = \{i : u_i \neq 0\}$ and $K_2 = \{i : v_i \neq 0\}$ be the sets indexing the non-zero components of the vectors $\mathbf{u}$ and $\mathbf{v}$, respectively. We assume that $\mathbf{u}$ and $\mathbf{v}$ are sparse, that is, $k_1 := |K_1| \ll n_1$ and $k_2 := |K_2| \ll n_2$. While the sets (K_1, K_2) are unknown, we assume that their cardinalities are known. Notice that the magnitude of the signal for all the coordinates in the bicluster $K_1 \times K_2$ is $\frac{\beta}{\sqrt{k_1 k_2}}$. The parameter β measures the strength of the signal, and is the key quantity we will be studying.

We focus on the case of a single bicluster that appears as an elevated sub-matrix of size $k_1 \times k_2$ with signal strength β embedded in a large $n_1 \times n_2$ data matrix with entries corrupted by additive Gaussian noise with variance σ^2. Under this model, the biclustering problem is formulated as the problem of estimating the sets K_1 and K_2, based on a single noisy observation $\mathbf{A}$ of the unknown signal matrix $\beta \mathbf{u} \mathbf{v}'$. Biclustering is most subtle when the matrix is large with several irrelevant variables, the entries are highly noisy, and the bicluster is small as defined by a sparse set of rows/columns. We provide a sharp characterization of tuples of $(\beta, n_1, n_2, k_1, k_2, \sigma^2)$ under which it is possible to recover the bicluster and study several common methods and establish the regimes under which they succeed.

We establish minimax lower and upper bounds for the following class of models. Let

$$\Theta(\beta_0, k_1, k_2) := \{(\beta, K_1, K_2) : \beta \geq \beta_0, |K_1| = k_1, K_1 \subset [n_1], |K_2| = k_2, K_2 \subset [n_2]\} \tag{2}$$

be a set of parameters. For a parameter $\theta \in \Theta$, let $\mathbb{P}_\theta$ denote the joint distribution of the entries of $\mathbf{A} = \{a_{ij}\}_{i \in [n_1], j \in [n_2]}$, whose density with respect to the Lebesgue measure is

$$\prod_{ij} \mathcal{N}(a_{ij}; \beta(k_1 k_2)^{-1/2} \mathbb{1}\{i \in K_1, j \in K_2\}, \sigma^2), \tag{3}$$

where the notation $\mathcal{N}(z; \mu, \sigma^2)$ denotes the distribution $p(z) \sim \mathcal{N}(\mu, \sigma^2)$ of a Gaussian random variable with mean μ and variance σ^2, and $\mathbb{1}$ denotes the indicator function.

We derive a lower bound that identifies tuples of $(\beta, n_1, n_2, k_1, k_2, \sigma^2)$ under which we can recover the true biclustering from a noisy high dimensional matrix. We show that a combinatorial procedure based on the scan statistic achieves the minimax optimal limits, however it is impractical as it requires enumerating all possible sub-matrices of a given size in a large matrix. We analyze the scalings (i.e. the relation between β and $(n_1, n_2, k_1, k_2, \sigma^2)$) under which some computationally tractable procedures for biclustering including element-wise thresholding, column/row average thresholding and sparse singular vector decomposition (SSVD) succeed with high probability.

We consider the detection of both small and large biclusters of weak activation, and show that at the minimax scaling the problem is surprisingly subtle (e.g., even detecting big clusters is quite hard).

In Table 1, we describe our main findings and compare the scalings under which the various algorithms succeed.

Algorithm	Combinatorial	Thresholding	Row/Column Averaging	Sparse SVD
SNR scaling	Minimax	Weak	Intermediate	Weak
Bicluster size	Any	Any	$(n_1^{1/2+\alpha} \times n_2^{1/2+\alpha}), \alpha \in (0, 1/2)$	Any
	Theorem 2	Theorem 3	Theorem 4	Theorem 5

Where the scalings are,

1. Minimax: $\beta \sim \sigma \max \left(\sqrt{k_1 \log(n_1 - k_1)}, \sqrt{k_2 \log(n_2 - k_2)} \right)$

2. Weak: $\beta \sim \sigma \max \left(\sqrt{k_1 k_2 \log(n_1 - k_1)}, \sqrt{k_1 k_2 \log(n_2 - k_2)} \right)$

3. Intermediate (for large clusters): $\beta \sim \sigma \max \left(\frac{\sqrt{k_1 k_2 \log(n_1 - k_1)}}{n_2^\alpha}, \frac{\sqrt{k_1 k_2 \log(n_2 - k_2)}}{n_1^\alpha} \right)$

Element-wise thresholding does not take advantage of any structure in the data matrix and hence does not achieve the minimax scaling for any bicluster size. If the clusters are big enough row/column averaging performs better than element-wise thresholding since it can take advantage of structure. We also study a convex relaxation for sparse SVD, based on the DSPCA algorithm proposed by [11] that encourages the singular vectors of the matrix to be supported over a sparse set of variables. However, despite the increasing popularity of this method, we show that it is only guaranteed to yield a sparse set of singular vectors when the SNR is quite high, equivalent to element-wise thresholding, and fails for stronger scalings of the SNR.

1.1 Related work

Due to its practical importance and difficulty biclustering has attracted considerable attention (for some recent surveys see [9, 27, 20, 22]). Broadly algorithms for biclustering can be categorized as either score-based searches, or spectral algorithms. Many of the proposed algorithms for identifying relevant clusters are based on heuristic searches whose goal is to identify large average sub-matrices or sub-matrices that are well fit by a two-way ANOVA model. Sun et. al. [26] provide some statistical backing for these exhaustive search procedures. In particular, they show how to construct a test via exhaustive search to distinguish when there is a small sub-matrix of weak activation from the "null" case when there is no bicluster.

The premise behind the spectral algorithms is that if there was a sub-matrix embedded in a large matrix, then this sub-matrix could be identified from the left and right singular vectors of $\mathbf{A}$. In the case when exactly one of $\mathbf{u}$ and $\mathbf{v}$ is random, the model (1) can be related to the spiked covariance model of [15]. In the case when $\mathbf{v}$ is random, the matrix $\mathbf{A}$ has independent columns and dependent rows. Therefore, $\mathbf{A}'\mathbf{A}$ is a spiked covariance matrix and it is possible to use the existing theoretical results on the first eigenvalue to characterize the left singular vector of $\mathbf{A}$. A lot of recent work has dealt with estimation of sparse eigenvectors of $\mathbf{A}'\mathbf{A}$, see for example [32, 16, 24, 31, 2]. For biclustering applications, the assumption that exactly one $\mathbf{u}$ or $\mathbf{v}$ is random, is not justifiable, therefore, theoretical results for the spiked covariance model do not translate directly. Singular vectors of the model (1) have been analyzed by [21], improving on earlier results of [6]. These results however are asymptotic and do not consider the case when $\mathbf{u}$ and $\mathbf{v}$ are sparse.

Our setup for the biclustering problem also falls in the framework of structured normal means multiple hypothesis testing problems, where for each entry in the matrix the hypotheses are that the entry has mean 0 versus an elevated mean. The presence of a bicluster (sub-matrix) however imposes structure on which elements are elevated concurrently. Recently, several papers have investigated the structured normal means setting for ordered domains. For example, [5] consider the detection of elevated intervals and other parametric structures along an ordered line or grid, [4] consider detection of elevated connected paths in tree and lattice topologies, [3] considers nonparametric cluster structures in a regular grid. In addition, [1] consider testing of different elevated structures in a general but known graph topology. Our setup for the biclustering problem requires identification of an elevated submatrix in an *unordered* matrix. At a high level, all these results suggest that it is possible to leverage the structure to improve the SNR threshold at which the hypothesis testing problem is

feasible. However, computationally efficient procedures that achieve the minimax SNR thresholds are only known for a few of these problems. Our results for biclustering have a similar flavor, in that the minimax threshold requires a combinatorial procedure whereas the computationally efficient procedures we investigate are often sub-optimal.

The rest of this paper is organized as follows. In Section 2, we provide a lower bound on the minimum signal strength needed for successfully identifying the bicluster. Section 3 presents a combinatorial procedure which achieves the lower bound and hence is minimax optimal. We investigate some computationally efficient procedures in Section 4. Simulation results are presented in Section 5 and we conclude in Section 6. All proofs are deferred to the Appendix.

2 Lower bound

In this section, we derive a lower bound for the problem of identifying the correct bicluster, indexed by K_1 and K_2, in model (1). In particular, we derive conditions on $(\beta, n_1, n_2, k_1, k_2, \sigma^2)$ under which any method is going to make an error when estimating the correct cluster. Intuitively, if either the signal-to-noise ratio β/σ or the cluster size is small, the minimum signal strength needed will be high since it is harder to distinguish the bicluster from the noise.

Theorem 1. *Let $\alpha \in (0, \frac{1}{8})$ and*

$$\beta_{\min} = \beta_{\min}(n_1, n_2, k_1, k_2, \sigma)$$

$$= \sigma\sqrt{\alpha}\max\left(\sqrt{k_1 \log(n_1 - k_1)}, \sqrt{k_2 \log(n_2 - k_2)}, \sqrt{\frac{k_1 k_2 \log(n_1 - k_1)(n_2 - k_1)}{k_1 + k_2 - 1}}\right).$$

(4)

Then for all $\beta_0 \leq \beta_{\min}$,

$$\inf_{\Phi} \sup_{\theta \in \Theta(\beta_0, k_1, k_2)} \mathbb{P}_\theta[\Phi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta))] \geq \frac{\sqrt{M}}{1 + \sqrt{M}}\left(1 - 2\alpha - \frac{2\alpha}{\log M}\right) \xrightarrow{n_1, n_2 \to \infty} 1 - 2\alpha,$$

(5)

where $M = \min(n_1 - k_1, n_2 - k_2)$, $\Theta(\beta_0, k_1, k_2)$ is given in (2) and the infimum is over all measurable maps $\Phi : \mathbb{R}^{n_1 \times n_2} \mapsto 2^{[n_1]} \times 2^{[n_2]}$.

The result can be interpreted in the following way: for any biclustering procedure Φ, if $\beta_0 \leq \beta_{\min}$, then there exists some element in the model class $\Theta(\beta_0, k_1, k_2)$ such that the probability of incorrectly identifying the sets K_1 and K_2 is bounded away from zero.

The proof is based on a standard technique described in Chapter 2.6 of [28]. We start by identifying a subset of parameter tuples that are hard to distinguish. Once a suitable finite set is identified, tools for establishing lower bounds on the error in multiple-hypothesis testing can be directly applied. These tools only require computing the Kullback-Leibler (KL) divergence between two distributions $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$, which in the case of model (1) are two multivariate normal distributions. These constructions and calculations are described in detail in the Appendix.

3 Minimax optimal combinatorial procedure

We now investigate a combinatorial procedure achieving the lower bound of Theorem 1, in the sense that, for any $\theta \in \Theta(\beta_{\min}, k_1, k_2)$, the probability of recovering the true bicluster (K_1, K_2) tends to one as n_1 and n_2 grow unbounded. This scan procedure consists in enumerating all possible pairs of subsets of the row and column indexes of size k_1 and k_2, respectively, and choosing the one whose corresponding submatrix has the largest overall sum. In detail, for an observed matrix $\mathbf{A}$ and two candidate subsets $\tilde{K}_1 \subset [n_1]$ and $\tilde{K}_2 \subset [n_2]$, we define the associated score $\mathcal{S}(\tilde{K}_1, \tilde{K}_2) := \sum_{i \in \tilde{K}_1} \sum_{j \in \tilde{K}_2} a_{ij}$. The estimated bicluster is the pair of subsets of sizes k_1 and k_2 achieving the highest score:

$$\Psi(\mathbf{A}) := \underset{(\tilde{K}_1, \tilde{K}_2)}{\operatorname{argmax}} \ \mathcal{S}(\tilde{K}_1, \tilde{K}_2) \quad \text{subject to} \quad |\tilde{K}_1| = k_1, \ |\tilde{K}_2| = k_2.$$

(6)

The following theorem determines the signal strength β needed for the decoder Ψ to find the true bicluster.

Theorem 2. *Let* $\mathbf{A} \sim \mathbb{P}_\theta$ *with* $\theta \in \Theta(\beta, k_1, k_2)$ *and assume that* $k_1 \leq n_1/2$ *and* $k_2 \leq n_2/2$. *If*

$$\beta \geq 4\sigma \max\left(\sqrt{k_1 \log(n_1 - k_1)}, \sqrt{k_2 \log(n_2 - k_2)}, \sqrt{\frac{2k_1 k_2 \log(n_1 - k_1)(n_2 - k_2)}{k_1 + k_2}}\right) \quad (7)$$

then $\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] \leq 2[(n_1 - k_1)^{-1} + (n_2 - k_2)^{-1}]$ *where* Ψ *is the decoder defined in* (6).

Comparing to the lower bound in Theorem 1, we observe that the combinatorial procedure using the decoder Ψ that looks for all possible clusters and chooses the one with largest score achieves the lower bound up to constants. Unfortunately, this procedure is not practical for data sets commonly encountered in practice, as it requires enumerating all $\binom{n_1}{k_1}\binom{n_2}{k_2}$ possible sub-matrices of size $k_1 \times k_2$. The combinatorial procedure requires the signal to be positive, but not necessarily constant throughout the bicluster. In fact it is easy to see that provided the average signal in the bicluster is larger than that stipulated by the theorem this procedure succeeds with high probability irrespective of how the signal is distributed across the bicluster. Finally, we remark that the estimation of the cluster is done under the assumption that k_1 and k_2 are known. Establishing minimax lower bounds and a procedure that adapts to unknown k_1 and k_2 is an open problem.

4 Computationally efficient biclustering procedures

In this section we investigate the performance of various procedures for biclustering, that, unlike the optimal scan statistic procedure studied in the previous section, are computationally tractable. For each of these procedures however, computational ease comes at the cost of suboptimal performance: recovery of the true bicluster is only possible if the β is much larger than the minimax signal strength of Theorem 1.

4.1 Element-wise thresholding

The simplest procedure that we analyze is based on element-wise thresholding. The bicluster is estimated as

$$\Psi_{\text{thr}}(\mathbf{A}, \tau) := \{(i, j) \in [n_1] \times [n_2] : |a_{ij}| \geq \tau\} \quad (8)$$

where $\tau > 0$ is a parameter. The following theorem characterizes the signal strength β required for the element-wise thresholding to succeed in recovering the bicluster.

Theorem 3. *Let* $\mathbf{A} \sim \mathbb{P}_\theta$ *with* $\theta \in \Theta(\beta, k_1, k_2)$ *and fix* $\delta > 0$. *Set the threshold* τ *as*

$$\tau = \sigma\sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}}.$$

If

$$\beta \geq \sqrt{k_1 k_2}\sigma\left(\sqrt{2 \log \frac{k_1 k_2}{\delta}} + \sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}}\right)$$

then $\mathbb{P}[\Psi_{\text{thr}}(\mathbf{A}, \tau) \neq K_1 \times K_2] = o(\delta/(k_1 k_2))$.

Comparing Theorem 3 with the lower bound in Theorem 1, we observe that the signal strength β needs to be $\mathcal{O}(\max(\sqrt{k_1}, \sqrt{k_2}))$ larger than the lower bound. This is not surprising, since the element-wise thresholding is not exploiting the structure of the problem, but is assuming that the large elements of the matrix $\mathbf{A}$ are positioned randomly. From the proof it is not hard to see that this upper bound is tight up to constants, i.e. if $\beta \leq c\sqrt{k_1 k_2}\sigma\left(\sqrt{2 \log \frac{k_1 k_2}{\delta}} + \sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}}\right)$ for a small enough constant c then thresholding will no longer recover the bicluster with probability at least $1 - \delta$. It is also worth noting that thresholding neither requires the signal in the bicluster to be constant nor positive provided it is larger in magnitude, at every entry, than the threshold specified in the theorem.

4.2 Row/Column averaging

Next, we analyze another a procedure based on column and row averaging. When the bicluster is large this procedure exploits the structure of the problem and outperforms the simple element-wise thresholding and the sparse SVD, which is discussed in the following section. The averaging procedure works only well if the bicluster is "large", as specified below, since otherwise the row or column average is dominated by the noise.

More precisely, the averaging procedure computes the average of each row and column of $\mathbf{A}$ and outputs the k_1 rows and k_2 columns with the largest average. Let $\{r_{r,i}\}_{i \in [n_1]}$ and $\{r_{c,j}\}_{j \in [n_2]}$ denote the positions of rows and columns when they are ordered according to row and column averages in descending order. The bicluster is estimated then as

$$\Psi_{\mathrm{avg}}(\mathbf{A}) := \{i \in [n_1] : r_{r,i} \leq k_1\} \times \{j \in [n_2] : r_{c,j} \leq k_2\}. \tag{9}$$

The following theorem characterizes the signal strength β required for the averaging procedure to succeed in recovering the bicluster.

Theorem 4. *Let $\mathbf{A} \sim \mathbb{P}_\theta$ with $\theta \in \Theta(\beta, k_1, k_2)$. If $k_1 = \Omega(n_1^{1/2+\alpha})$ and $k_2 = \Omega(n_2^{1/2+\alpha})$, where $\alpha \in (0, 1/2)$ is a constant and,*

$$\beta \geq 4\sigma \max\left(\frac{\sqrt{k_1 k_2 \log(n_1 - k_1)}}{n_2^\alpha}, \frac{\sqrt{k_1 k_2 \log(n_2 - k_2)}}{n_1^\alpha} \right)$$

then $\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] \leq [n_1^{-1} + n_2^{-1}]$.

Comparing to Theorem 3, we observe that the averaging requires lower signal strength than the element-wise thresholding when the bicluster is large, that is, $k_1 = \Omega(\sqrt{n_1})$ and $k_2 = \Omega(\sqrt{n_2})$. Unless both $k_1 = \mathcal{O}(n_1)$ and $k_2 = \mathcal{O}(n_2)$, the procedure does not achieve the lower bound of Theorem 1, however, the procedure is simple and computationally efficient. It is also not hard to show that this theorem is sharp in its characterization of the averaging procedure. Further, unlike thresholding, averaging requires the signal to be positive in the bicluster.

It is interesting to note that a large bicluster can also be identified without assuming the normality of the noise matrix $\mathbf{\Delta}$. This non-parametric extension is based on a simple sign-test, and the details are provided in Appendix.

4.3 Sparse singular value decomposition (SSVD)

An alternate way to estimate K_1 and K_2 would be based on the singular value decomposition (SVD), i.e. finding $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ that maximize $\langle \tilde{\mathbf{u}}, \mathbf{A}\tilde{\mathbf{v}} \rangle$, and then threshold the elements of $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$. Unfortunately, such a method would perform poorly when the signal β is weak and the dimensionality is high, since, due to the accumulation of noise, $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are poor estimates of $\mathbf{u}$ and $\mathbf{v}$ and and do not exploit the fact that $\mathbf{u}$ and $\mathbf{v}$ are sparse.

In fact, it is now well understood [8] that SVD is strongly inconsistent when the signal strength is weak, i.e. $\angle(\tilde{\mathbf{u}}, \mathbf{u}) \to \pi/2$ (and similarly for $\mathbf{v}$) almost surely. See [26] for a clear exposition and discussion of this inconsistency in the SVD setting.

To properly exploit the sparsity in the singular vectors, it seems natural to impose a cardinality constraint to obtain a sparse singular vector decomposition (SSVD):

$$\max_{\mathbf{u} \in \mathbf{S}^{n_1-1}, \mathbf{v} \in \mathbf{S}^{n_2-1}} \langle \mathbf{u}, \mathbf{A}\mathbf{v} \rangle \quad \text{subject to} \quad ||\mathbf{u}||_0 \leq k_1, ||\mathbf{v}||_0 \leq k_2,$$

which can be further rewritten as

$$\max_{\mathbf{Z} \in \mathbb{R}^{n_2 \times n_1}} \mathrm{tr}\,\mathbf{AZ} \quad \text{subject to} \quad \mathbf{Z} = \mathbf{vu}', ||\mathbf{u}||_2 = 1, ||\mathbf{v}||_2 = 1, ||\mathbf{u}||_0 \leq k_1, ||\mathbf{v}||_0 \leq k_2. \tag{10}$$

The above problem is non-convex and computationally intractable.

Inspired by the convex relaxation methods for sparse principal component analysis proposed by [11], we consider the following relaxation the SSVD:

$$\max_{\mathbf{X} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)}} \mathrm{tr}\,\mathbf{AX}^{21} - \lambda \mathbf{1}'|\mathbf{X}^{21}|\mathbf{1} \quad \text{subject to} \quad \mathbf{X} \succeq \mathbf{0}, \mathrm{tr}\,\mathbf{X}^{11} = 1, \mathrm{tr}\,\mathbf{X}^{22} = 1, \tag{11}$$

where $\mathbf{X}$ is the block matrix

$$\begin{bmatrix} \mathbf{X}^{11} & \mathbf{X}^{12} \\ \mathbf{X}^{21} & \mathbf{X}^{22} \end{bmatrix}$$

with the block $\mathbf{X}^{21}$ corresponding to $\mathbf{Z}$ in (10). If the optimal solution $\widehat{\mathbf{X}}$ is of rank 1, then, necessarily, $\widehat{\mathbf{X}} = \binom{\widehat{\mathbf{u}}}{\widehat{\mathbf{v}}}(\widehat{\mathbf{u}}' \ \widehat{\mathbf{v}}')$. Based on the sparse singular vectors $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$, we estimate the bicluster as

$$\widehat{K}_1 = \{j \in [n_1] \ : \ \widehat{u}_j \neq 0\} \qquad \text{and} \qquad \widehat{K}_2 = \{j \in [n_2] \ : \ \widehat{v}_j \neq 0\}. \tag{12}$$

The user defined parameter λ controls the sparsity of the solution $\widehat{\mathbf{X}}^{21}$, and, therefore, provided the solution is of rank one, it also controls the sparsity of the vectors $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ and of the estimated bicluster.

The following theorem provides *sufficient* conditions for the solution $\widehat{\mathbf{X}}$ to be rank one and to recover the bicluster.

Theorem 5. *Consider the model in* (1). *Assume* $k_1 \asymp k_2$ *and* $k_1 \leq n_1/2$ *and* $k_2 \leq n_2/2$. *If*

$$\beta \geq 2\sigma\sqrt{k_1 k_2 \log(n_1 - k_1)(n_2 - k_2)} \tag{13}$$

then the solution $\widehat{\mathbf{X}}$ *of the optimization problem in* (11) *with* $\lambda = \frac{\beta}{2\sqrt{k_1 k_2}}$ *is of rank 1 with probability* $1 - \mathcal{O}(k_1^{-1})$. *Furthermore, we have that* $(\widehat{K}_1, \widehat{K}_2) = (K_1, K_2)$ *with probability* $1 - \mathcal{O}(k_1^{-1})$.

It is worth noting that SSVD correctly recovers *signed* vectors $\widehat{u}$ and $\widehat{v}$ under this signal strength. In particular, the procedure works even if the u and v in Equation 1 are signed.

The following theorem establishes *necessary* conditions for the SSVD to have a rank 1 solution that correctly identifies the bicluster.

Theorem 6. *Consider the model in* (1). *Fix* $c \in (0, 1/2)$. *Assume that* $k_1 \asymp k_2$ *and* $k_1 = o(n^{1/2-c})$ *and* $k_2 = o(n_2^{1/2-c})$. *If*

$$\beta \leq 2\sigma\sqrt{c k_1 k_2 \log \max(n_1 - k_1, n_2 - k_2)}, \tag{14}$$

with $\lambda = \frac{\beta}{2\sqrt{k_1 k_2}}$ *then the optimization problem* (11) *does not have a rank 1 solution that correctly recovers the sparsity pattern with probability at least* $1 - \mathcal{O}(\exp(-(\sqrt{k_1} + \sqrt{k_2})^2))$ *for sufficiently large* n_1 *and* n_2.

From Theorem 6 observe that the sufficient conditions of Theorem 5 are sharp. In particular, the two theorems establish that the SSVD does not establish the lower bound given in Theorem 1. The signal strength needs to be of the same order as for the element-wise thresholding, which is somewhat surprising since from the formulation of the SSVD optimization problem it seems that the procedure uses the structure of the problem. From numerical simulations in Section 5 we observe that although SSVD requires the same scaling as thresholding, it consistently performs slightly better at a fixed signal strength.

5   Simulation results

We test the performance of the three computationally efficient procedures on synthetic data: thresholding, averaging and sparse SVD. For sparse SVD we use an implementation posted online by [11]. We generate data from (1) with $n = n_1 = n_2$, $k = k_1 = k_2$, $\sigma^2 = 1$ and $\mathbf{u} = \mathbf{v} \propto (\mathbf{1}_k', \mathbf{0}_{n-k}')'$. For each algorithm we plot the Hamming fraction (i.e. the Hamming distance between $\mathbf{s}_{\widehat{\mathbf{u}}}$ and $\mathbf{s}_{\mathbf{u}}$ rescaled to be between 0 and 1) against the rescaled sample size. In each case we average the results over 50 runs.

For thresholding and sparse SVD the rescaled scaling (x-axis) is $\frac{\beta}{k\sqrt{\log(n-k)}}$ and for averaging the rescaled scaling (x-axis) is $\frac{\beta n^{\alpha}}{k\sqrt{\log(n-k)}}$. We observe that there is a sharp threshold between success and failure of the algorithms, and the curves show good agreement with our theory.

The vertical line shows the point after which successful recovery happens for all values of n. We can make a direct comparison between thresholding and sparse SVD (since the curves are identically rescaled) to see that at least empirically sparse SVD succeeds at a smaller scaling constant than thresholding even though their asymptotic rates are identical.
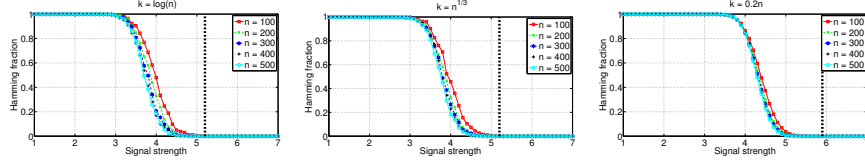
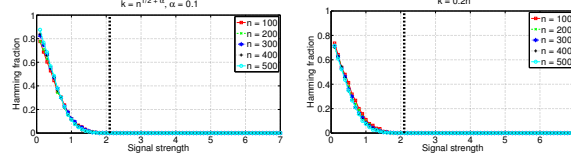Figure 1: Thresholding: Hamming fraction versus rescaled signal strength.

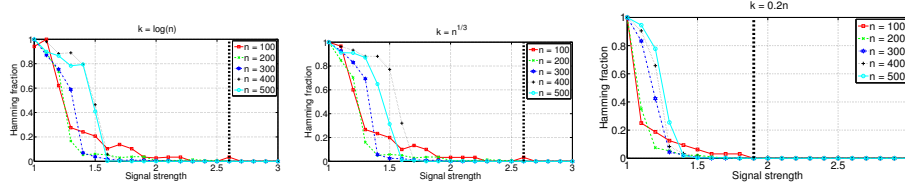Figure 2: Averaging: Hamming fraction versus rescaled signal strength.

Figure 3: Sparse SVD: Hamming fraction versus rescaled signal strength.

6 Discussion

In this paper, we analyze biclustering using a simple statistical model (1), where a sparse rank one matrix is perturbed with noise. Using this model, we have characterized the minimal signal strength below which no procedure can succeed in recovering the bicluster. This lower bound can be matched using an exhaustive search technique. However, it is still an open problem to find a computationally efficient procedure that is minimax optimal.

Amini et. al. [2] analyze the convex relaxation procedure proposed in [11] for high-dimensional sparse PCA. Under the minimax scaling for this problem they show that provided a rank-1 solution exists it has the desired sparsity pattern (they were however not able to show that a rank-1 solution exists with high probability). Somewhat surprisingly, we show that in the SVD case a rank-1 solution with the desired sparsity pattern *does not* exist with high probability. The two settings however are not identical since the noise in the spiked covariance model is Wishart rather than Gaussian, and has correlated entries. It would be interesting to analyze whether our negative result has similar implications for the sparse PCA setting.

The focus of our paper has been on a model with one cluster, which although simple, provides several interesting theoretical insights. In practice, data often contains multiple clusters which need to be estimated. Many existing algorithms (see e.g. [17] and [18]) try to estimate multiple clusters and it would be useful to analyze these theoretically.

Furthermore, the algorithms that we have analyzed assume knowledge of the size of the cluster, which is used to select the tuning parameters. It is a challenging problem of great practical relevance to find data driven methods to select these tuning parameters.

7 Acknowledgments

References

[1] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi. On combinatorial testing problems. *Ann. Statist.*, 38(5):3063–3092, 2010.

[2] A.A. Amini and M.J. Wainwright. High-Dimensional Analysis Of Semidefinite Relaxations For Sparse Principal Components. *The Annals of Statistics*, 37(5B):2877–2921, 2009.

[3] Ery Arias-Castro, Emmanuel J. Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *Ann. Stat.*, 39(1):278–304, 2011.

[4] Ery Arias-Castro, Emmanuel J. Candès, Hannes Helgason, and Ofer Zeitouni. Searching for a trail of evidence in a maze. *Ann. Statist.*, 36(4):1726–1757, 2008.

[5] Ery Arias-Castro, David L. Donoho, and Xiaoming Huo. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Statist.*, 34(1):326–349, 2006.

[6] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):pp. 135–171, 2003.

[7] Ulrich Bayer, Paolo Milani Comparetti, Clemens Hlauscheck, Christopher Kruegel, and Engin Kirda. Scalable, Behavior-Based Malware Clustering. In *16th Symposium on Network and Distributed System Security (NDSS)*, 2009.

[8] F. Benaych-Georges and R. Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *ArXiv e-prints*, March 2011.

[9] S. Busygin, O. Prokopyev, and P.M. Pardalos. Biclustering in data mining. *Computers & Operations Research*, 35(9):2964–2987, 2008.

[10] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *CoRR*, abs/0912.3599, 2009.

[11] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49:434–448, 2007.

[12] K.R. Davidson and S.J. Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1:317–366, 2001.

[13] R. Fletcher. Semi-definite matrix constraints in optimization. *SIAM Journal on Control and Optimization*, 23:493, 1985.

[14] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):pp. 123–129, 1972.

[15] I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[16] I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[17] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica sinica*, 12:61–86, 2002.

[18] Mihee Lee, Haipeng Shen, Jianhua Z. Huang, and J. S. Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010.

[19] Jinze Liu and Wei Wang. Op-cluster: Clustering by tendency in high dimensional space. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pages 187–, Washington, DC, USA, 2003. IEEE Computer Society.

[20] S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on computational Biology and Bioinformatics*, pages 24–45, 2004.

[21] A. Onatski. Asymptotics of the principal components estimator of large factor models with weak factors. *Economics Department, Columbia University*, 2009.

[22] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

[23] R.T. Rockafellar. *The theory of subgradients and its applications to problems of optimization. Convex and nonconvex functions*. Heldermann, 1981.

[24] H. Shen and J.Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008.

[25] GW Stewart. Perturbation theory for the singular value decomposition. *Computer Science Technical Report Series; Vol. CS-TR-2539*, page 13, 1990.

[26] X. Sun and A. B. Nobel. On the maximal size of Large-Average and ANOVA-fit Submatrices in a Gaussian Random Matrix. *ArXiv e-prints*, September 2010.

[27] A. Tanay, R. Sharan, and R. Shamir. Biclustering algorithms: A survey. *Handbook of computational molecular biology*, 2004.

[28] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

[29] Lyle Ungar and Dean P. Foster. A formal statistical approach to collaborative filtering. In *CONALD*, 98.

[30] S. Wang, R. R. Gutell, and D. P. Miranker. Biclustering as a method for RNA local multiple sequence alignment. *Bioinformatics*, 23:3289–3296, Dec 2007.

[31] D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515, 2009.

[32] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

8 Appendix: Proofs of the main results

This section collects proofs of the main results stated in Sections 2, 3 and 4.

8.1 Proof of Theorem 1

We use a standard technique based on multiple hypothesis testing to obtain a lower bound on the minimal signal strength (see Section 2.6. in [28]). Without loss of generality, we assume $\sigma = 1$. Set $K_1 = [k_1]$ and $K_2 = [k_2]$, and let $\tau_0 = \beta(k_1 k_2)^{-1/2}$, so that the joint density of $\mathbf{A}$ is

$$\prod_{ij} \mathcal{N}(a_{ij}; \tau_0 \, \mathbb{I}\{i \in K_1, j \in K_2\}, 1).$$

To lower bound the probability of error, we use the following relationship

$$\inf_{\Psi} \sup_{\theta \in \Theta} \mathbb{P}_\theta(\Psi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta))) \geq \inf_{\Psi} \max_{\theta \in \{\theta_0, \dots, \theta_M\}} \mathbb{P}_\theta(\Psi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta)))$$

where $\{\theta_0, \theta_1, \dots, \theta_M\}$ is a carefully chosen subset of Θ. Specifically, we select $\theta_0 = (\beta, K_1, K_2)$ and we choose the remaining points $\{\theta_1, \dots, \theta_M\}$, with $M = n_2 - k_2$, so that

$$\theta_{j-k_2} = (\beta, K_1, K_2^{(j)}), \quad j = k_2 + 1, \dots, n_2,$$

where $K_2^{(j)} := [k_2 - 1] \cup \{j\}$. For a $\theta \in \Theta$, below we denote with $(K_1(\theta), K_2(\theta))$ the associated bicluster.

Let $\phi(u)$ denote the density function of $\mathcal{N}(0, 1)$ with respect to the Lebesgue measure. With this, we can compute the Kullback-Leibler divergence between $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_j}$:

$$\begin{aligned}
D(\mathbb{P}_{\theta_0} | \mathbb{P}_{\theta_j}) &= \int \log \frac{d\mathbb{P}_{\theta_0}}{d\mathbb{P}_{\theta_j}} d\mathbb{P}_{\theta_0} \\
&= \sum_{i \in K_1} \int \log \frac{\phi(u_{ik_2} - \tau_0)}{\phi(u_{ik_2})} \phi(u_{ik_2} - \tau_0) du_{ik_2} \\
&\quad + \sum_{i \in K_1} \int \log \frac{\phi(u_{ij})}{\phi(u_{ij} - \tau_0)} \phi(u_{ij}) du_{ij} \\
&= \sum_{i \in K_1} \int (u_{ik_2} \tau_0 - \frac{\tau_0^2}{2}) \phi(u_{ik_2} - \tau_0) du_{ik_2} \\
&\quad + \sum_{i \in K_1} \int (\frac{\tau_0^2}{2} - u_{ij} \tau_0) \phi(u_{ij}) du_{ij} \\
&= \sum_{i \in K_1} \int u_{ik_2} \tau_0 \phi(u_{ik_2} - \tau_0) du_{ik_2} \\
&= k_1 \tau_0^2.
\end{aligned} \tag{15}$$

Now it follows from Theorem 2.5 in [28] that, if

$$\tau_0 \leq \sqrt{\frac{\alpha \log(n_2 - k_2)}{k_1}},$$

then

$$\inf_{\Psi} \max_{\theta \in \{\theta_0, \dots, \theta_M\}} \mathbb{P}_\theta(\Psi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta))) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \frac{2\alpha}{\log M}\right) \xrightarrow{n_1, n_2 \to \infty} 1 - 2\alpha.$$

We chose the subset $\{\theta_1, \dots, \theta_M\}$ by fixing the set K_1 and alternating the last element of the set K_2. Alternatively, we can fix K_2 and change the last element of the set K_1 or alternate both K_1 and

10

K_2. Repeating the argument above for these cases, we have that the probability of making an error is bounded away from zero if

$$\tau_0 \leq \max\left(\sqrt{\frac{\alpha \log(n_2 - k_2)}{k_1}}, \sqrt{\frac{\alpha \log(n_1 - k_1)}{k_2}}, \sqrt{\frac{\alpha \log(n_1 - k_1)(n_2 - k_1)}{k_1 + k_2 - 1}}\right), \tag{16}$$

which completes the proof.

8.2 Proof of Theorem 2

Without loss of generality, we assume that the noise variance $\sigma = 1$ and the true unknown sets $K_1 = [k_1]$ and $K_2 = [k_2]$. Define

$$F(\tilde{K}_1, \tilde{K}_2) := \sum_{i \in K_1} \sum_{j \in K_2} \mathbf{A}_{ij} - \sum_{i \in \tilde{K}_1} \sum_{j \in \tilde{K}_2} \mathbf{A}_{ij} \tag{17}$$

and note that an error is made if $F(\tilde{K}_1, \tilde{K}_2) < 0$, so that

$$\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] = \mathbb{P}[\cup_{\tilde{K}_1, \tilde{K}_2}\{F(\tilde{K}_1, \tilde{K}_2) < 0\}].$$

Observe that $F(\tilde{K}_1, \tilde{K}_2)$ depends only on the amount of overlap between $K_1 \times K_2$ and $\tilde{K}_1 \times \tilde{K}_2$. In particular, we have that

$$F(\tilde{K}_1, \tilde{K}_2) = F(d) \overset{d}{=} \mathcal{N}(d\beta(k_1 k_2)^{-1/2}, 2d\sigma^2) \tag{18}$$

where $d = k_1 k_2 - |K_1 \cap \tilde{K}_1||K_2 \cap \tilde{K}_2|$. Therefore, using the union bound, we have that

$$\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] \leq \sum_{i=0}^{k_1} C_{k_1}^i C_{n_1-k_1}^{k_1-i} \sum_{j=0}^{k_2} C_{k_2}^j C_{n_2-k_2}^{k_2-j} \mathbb{P}[F(k_1 k_2 - ij) < 0],$$

where, for readability, we have adopted the notation $C_n^i = \binom{n}{i}$.

Let $\tau_0 = \beta(k_1 k_2)^{-1/2}$. Using (18),

$$\mathbb{P}(\Psi(\mathbf{A}) \neq (K_1, K_2)) \leq \sum_{i=0}^{k_1} C_{k_1}^i C_{n_1-k_1}^{k_1-i} \sum_{j=0}^{k_2} C_{k_2}^j C_{n_2-k_2}^{k_2-j} \mathbb{P}(F(k_1 k_2 - ij) < 0)$$

$$= \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} p_{ij} - p_{k_1 k_2}$$

with

$$p_{ij} = C_{k_1}^i C_{n_1-k_1}^{k_1-i} C_{k_2}^j C_{n_2-k_2}^{k_2-j} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2})$$

and $\bar{\Phi}(\cdot)$ is the survival function of $\mathcal{N}(0,1)$. Therefore, $\mathbb{P}(\Psi(\mathbf{A}) \neq (K_1, K_2))$ can be bounded by

$$\underbrace{(k_1 - 1)(k_2 - 1) \max_{\substack{i=0,\ldots,k_1-1 \\ j=0,\ldots,k_2-1}} p_{ij}}_{T_1} + \underbrace{(k_1 - 1) \max_{i=0,\ldots,k_1-1} p_{ik_2}}_{T_2} + \underbrace{(k_2 - 1) \max_{j=0,\ldots,k_2-1} p_{k_1 j}}_{T_3}.$$

We'll show how to handle T_1, while T_2 and T_3 can be handled in an similar way.

$$
\begin{aligned}
T_1 &= (k_1 - 1)(k_2 - 1) \max_{\substack{i=0,\ldots,k_1-1 \\ j=0,\ldots,k_2-1}} C_{k_1}^i C_{n_1-k_1}^{k_1-i} C_{k_2}^j C_{n_2-k_2}^{k_2-j} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2}) \\
&\leq (k_1 - 1)(k_2 - 1) \max_{\substack{i=0,\ldots,k_1-1 \\ j=0,\ldots,k_2-1}} (n_1 - k_1)^{2(k_1-i)} (n_2 - k_2)^{2(k_2-j)} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2}) \\
&\leq \max_{\substack{i=0,\ldots,k_1-1 \\ j=0,\ldots,k_2-1}} (n_1 - k_1)^{3(k_1-i)} (n_2 - k_2)^{3(k_2-j)} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2}) \\
&\leq \max_{\substack{i=0,\ldots,k_1-1 \\ j=0,\ldots,k_2-1}} (n_1 - k_1)^{3(k_1-i)} (n_2 - k_2)^{3(k_2-j)} \exp\left\{-\frac{\tau_0^2}{4}\left(k_1 k_2 - \frac{ik_2}{2} - \frac{jk_1}{2}\right)\right\}
\end{aligned}
$$

It is easy to see that the maximum is achieved at $i = k_1 - 1$ and $j = k_2 - 1$, which gives

$$T_1 \leq (n_1 - k_1)^3 (n_2 - k_2)^3 \exp\left(-\frac{\tau_0^2 (k_1 + k_2)}{8}\right).$$

Using the same reasoning

$$T_2 \leq (n_1 - k_1)^3 \exp\left(-\frac{\tau_0^2 k_2}{4}\right) \quad \text{and} \quad T_3 \leq (n_2 - k_2)^3 \exp\left(-\frac{\tau_0^2 k_1}{4}\right).$$

Probability of making an error can be bounded as $\mathbb{P}(\Psi(\mathbf{A}) \neq (K_1, K_2)) \leq T_1 + T_2 + T_3$, which concludes the proof.

8.3 Proof of Theorem 3

The proof follows from an applications of the union bound and the tail bound for the standard normal random variable given in (25). We have that

$$\min_{(i,j) \in K_1 \times K_2} |a_{ij}| \geq (k_1 k_2)^{-1/2} \beta - \max_{(i,j) \in K_1 \times K_2} |\Delta_{ij}| \geq (k_1 k_2)^{-1/2} \beta - \sigma \sqrt{2 \log \frac{k_1 k_2}{\delta}}$$

with probability $1 - 2\delta_1/(\sqrt{4\pi \log(1/\delta_1)})$ where $\delta_1 = \delta/(k_1 k_2)$. Similarly,

$$\max_{(i,j) \notin K_1 \times K_2} |a_{ij}| = \max_{(i,j) \notin K_1 \times K_2} |\Delta_{ij}| \leq \sigma \sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}}$$

with probability $1 - 2\delta_2/\sqrt{4\pi \log(1/\delta_2)}$ where $\delta_2 = \delta/|\{(i,j) \notin K_1 \times K_2\}|$. Combining the last two displays, the theorem follows.

8.4 Proof of Theorem 4

First consider identifying the rows. The sum of the elements of each row without activation is a draw from $\mathcal{N}(0, n_2\sigma^2)$ and there are $(n_1 - k_1)$ of these, while the sum of the elements of each row with activation is a draw from $\mathcal{N}(4\sigma \max\left(\sqrt{n_2 \log(n_1)}, \sqrt{n_2 \log(n_2)}\left(\frac{n_2}{n_1}\right)^\alpha\right), n_2\sigma^2)$, and there are k_1 of these.

Consider the probability that all the rows without activation have sum strictly less than $2\sigma\sqrt{n_2 \log(n_1)}$, and those with activation have sum strictly greater than the same quantity. If this condition is satisfied then selecting the k_1 rows with highest sum produces no errors. It is also easy to see that to upper bound the probability of error it suffices to show that the probability of error is small if the activation rows were drawn from $\mathcal{N}(4\sigma\sqrt{n_2 \log(n_1)}, n_2\sigma^2)$.

The result follows from applying a standard Gaussian tail bound, followed by a union bound, i.e.

$$\mathbb{P}(X - \mu > t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

therefore, noting the symmetry we can bound

$$\mathbb{P}(\text{error}) \leq n_1 \exp\left(-\frac{4\sigma^2 n_2 \log(n_1)}{2n_2\sigma^2}\right) = n_1(n_1)^{-2} = \delta_1$$

A similar argument shows that we can bound δ_2, the probability of making an error in identifying the columns. The result follows.

8.5 Proof of Theorem 5

We prove the theorem using a constructive procedure. Our arguments are adapted from the arguments used in the proof of Theorem 2 in [2]. We construct a rank one solution $\widehat{\mathbf{X}}$ that is a global

solution of the problem in (11). Using Theorem 22, which states the first order conditions for a global optimum, we have that

$$-\begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}' & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}}' & \mathbf{0} \end{pmatrix} + (\widehat{\pi}_1 - \widehat{\pi}_2)\mathbf{I}_{n_1+n_2} = \widehat{\mathbf{K}}, \qquad (19)$$

where $\widehat{\mathbf{S}} \in \partial\|\widehat{\mathbf{X}}\|_1$ is an element of the subgradient of the element-wise ℓ_1 norm evaluated at $\widehat{\mathbf{X}}$, $\widehat{\pi}_1$ and $\widehat{\pi}_2$ are Lagrange multipliers associated with the constraint $\operatorname{tr}\widehat{\mathbf{X}} = 2$, and $\widehat{\mathbf{K}}$ is an element of the normal cone to $\mathcal{S}_+^n$ evaluated at $\widehat{\mathbf{X}}$. For $\widehat{\mathbf{S}}$, we have that $\max_{ij}|\widehat{S}_{ij}| \le 1$ and $\operatorname{tr}\widehat{\mathbf{S}}'\mathbf{X}^{12} = \mathbf{1}'|\mathbf{X}|\mathbf{1}$. From Eq (30), we have that $\widehat{\mathbf{K}} = -\widehat{\mathbf{Z}}^\perp\mathbf{B}\widehat{\mathbf{Z}}^\perp$ where columns of $\widehat{\mathbf{Z}}^\perp$ form orthonormal basis for the null space of $\widehat{\mathbf{X}}$ and $\mathbf{B} \in \mathcal{S}_+^n$. See §12 for more details.

Suppose that the matrix $\widehat{\mathbf{X}}$ is rank one and that the sparsity pattern of $\widehat{\mathbf{X}}^{12}$ correctly recovers K_1 and K_2. Then we have that $\widehat{\mathbf{S}}_{K_1K_2} = \mathbf{s}_{\widehat{\mathbf{u}}}\mathbf{s}_{\widehat{\mathbf{v}}}'$ where $\mathbf{s}_{\widehat{\mathbf{u}}} = \operatorname{sign}(\widehat{\mathbf{u}}_{K_1})$ and $\mathbf{s}_{\widehat{\mathbf{v}}} = \operatorname{sign}(\widehat{\mathbf{v}}_{K_2})$. Furthermore, $\widehat{\mathbf{X}}^{12}_{K_1K_2} = \widehat{\mathbf{u}}_{K_1}\widehat{\mathbf{v}}'_{K_2}$ where $\widehat{\mathbf{u}}_{K_1}$ is a left singular vector and $\widehat{\mathbf{v}}_{K_2}$ is a right singular vector of $\mathbf{A}_{K_1K_2} - \lambda\widehat{\mathbf{S}}_{K_1K_2}$ associated with the largest singular vector. In fact, the following Lemma will show that $\widehat{\mathbf{u}}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}$ are left and right singular vectors of $\mathbf{A}_{K_1K_2} - \lambda\mathbf{s}_{\mathbf{u}}\mathbf{s}_{\mathbf{v}}'$ where $\mathbf{s}_{\mathbf{u}} = \operatorname{sign}(\mathbf{u}_{K_1})$ and $\mathbf{s}_{\mathbf{v}} = \operatorname{sign}(\mathbf{v}_{K_2})$. That is, $\mathbf{s}_{\widehat{\mathbf{u}}}$ and $\mathbf{s}_{\widehat{\mathbf{v}}}$ recover signs of $\mathbf{s}_{\mathbf{u}}$ and $\mathbf{s}_{\mathbf{v}}$. Note that singular vectors are uniquely defined only up to a rotation, therefore, we use a convention that the first non-zero coordinate of a left singular vector is positive.

Let $\mathbf{M} = \mathbf{A}_{K_1K_2} - \lambda\operatorname{sign}(\mathbf{u}_{K_1})\operatorname{sign}(\mathbf{v}_{K_2})'$ and let $\alpha = \beta/2$. Since $\lambda = \frac{\beta}{2\sqrt{k_1k_2}}$, we have that $\mathbf{M} = \alpha\mathbf{u}_{K_1}\mathbf{v}'_{K_2} + \mathbf{\Delta}_{K_1K_2}$. Let $\widehat{\alpha} = \sigma_1(\mathbf{M})$ be the largest singular value of $\mathbf{M}$.

Lemma 7. *Under the conditions of Theorem 5, we have that*

$$\|\widehat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log k_1}{k_1k_2\log(n_1-k_2)(n_2-k_2)}}\right)$$

and

$$\|\widehat{\mathbf{v}}_{K_2} - \mathbf{v}_{K_2}\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log k_2}{k_1k_2\log(n_1-k_2)(n_2-k_2)}}\right)$$

with probability $1 - \mathcal{O}(k_1^{-1})$.

Under the assumptions of Theorem 5 $\|\widehat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_\infty = o(1/\sqrt{k_1})$ and $\|\widehat{\mathbf{v}}_{K_2} - \mathbf{v}_{K_2}\|_\infty = o(1/\sqrt{k_2})$ as $n_1, n_2 \to \infty$, which shows that $\mathbf{s}_{\widehat{\mathbf{u}}}$ and $\mathbf{s}_{\widehat{\mathbf{v}}}$ recover signs of $\mathbf{s}_{\mathbf{u}}$ and $\mathbf{s}_{\mathbf{v}}$.

Next, we set elements of $\widehat{\mathbf{S}}_{K_1^C K_2}$ and $\widehat{\mathbf{S}}_{K_1 K_2^C}$ such that $(\widehat{\mathbf{u}}'_{K_1}, \mathbf{0}')'$ and $(\widehat{\mathbf{v}}'_{K_2}, \mathbf{0}')'$ are singular vectors of $\mathbf{A} - \lambda\widehat{\mathbf{S}}$. Note that for these two singular vectors the choice of $\widehat{\mathbf{S}}_{K_1^C K_2^C}$ is irrelevant. Let $\widehat{\mathbf{S}}_{K_1^C K_2} = \lambda^{-1}\Delta_{K_1^C K_2}$ and $\widehat{\mathbf{S}}_{K_1 K_2^C} = \lambda^{-1}\Delta_{K_1 K_2^C}$. Using a normal tail bound (25) and the union bound

$$\|\widehat{\mathbf{S}}_{K_1^C K_2}\|_\infty \le \frac{4\sigma\sqrt{k_1k_2\log[(n_1-k_1)k_2]}}{\beta} \quad \text{and} \quad \|\widehat{\mathbf{S}}_{K_1 K_2^C}\|_\infty \le \frac{4\sigma\sqrt{k_1k_2\log[(n_2-k_2)k_1]}}{\beta}$$

with probability $1 - \mathcal{O}[(n_1-k_1)^{-1}k_2^{-1}]$. Under the assumptions of the theorem we have that $\|\widehat{\mathbf{S}}_{K_1 K_2^C}\|_\infty < 1$ and $\|\widehat{\mathbf{S}}_{K_1 K_2^C}\|_\infty < 1$.

Let $\widehat{\mathbf{x}} = (\widehat{\mathbf{u}}'_{K_1}, \mathbf{0}', \widehat{\mathbf{v}}'_{K_2}, \mathbf{0}')'$, so that $\widehat{\mathbf{X}} = \widehat{\mathbf{x}}\widehat{\mathbf{x}}'$. We have established so far that $\widehat{\mathbf{x}}$ is an eigenvector of

$$-\begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}' & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}}' & \mathbf{0} \end{pmatrix}.$$

Therefore, multiplying Eq. (19) by $\widehat{\mathbf{x}}$ from right and taking a dot product with $\widehat{\mathbf{x}}$ we have that $\widehat{\alpha} = \widehat{\pi}_1 - \widehat{\pi}_2$. Finally, we need to set $\widehat{\mathbf{S}}_{K_1^C K_2^C}$ such that (19) holds. Set $\widehat{\mathbf{K}}$ to the left hand side of (19), then we need to show that $\widehat{\mathbf{K}} \succeq \mathbf{0}$. By construction of $\widehat{\mathbf{X}}$, we have that $\widehat{\mathbf{K}}_{(K_1K_2)(K_1K_2)} \succeq \mathbf{0}$. Therefore, we only need to show that $\widehat{\mathbf{K}}_{(K_1^C K_2^C)(K_1^C K_2^C)} \succeq$

$\widehat{\mathbf{K}}_{(K_1^C K_2^C)(K_1 K_2)} (\widehat{\mathbf{K}}_{(K_1 K_2)(K_1 K_2)})^\dagger \widehat{\mathbf{K}}_{(K_1 K_2)(K_1^C K_2^C)}$. With the current choice of $\widehat{\mathbf{S}}_{K_1^C K_2}$ and $\widehat{\mathbf{S}}_{K_1 K_2^C}$, we can choose $\widehat{\mathbf{S}}_{K_1^C K_2^C} = \lambda^{-1} \mathbf{\Delta}_{K_1^C K_2^C}$ to satisfy (19). From (25) and the union bound

$$\|\widehat{\mathbf{S}}_{K_1^C K_2^C}\|_\infty \leq \frac{4\sigma \sqrt{k_1 k_2 \log[(n_1 - k_1)(n_2 - k_2)]}}{\beta}$$

with probability $1 - \mathcal{O}((n_1 - k_1)^{-1}(n_1 - k_1)^{-1})$. Under the assumptions of the theorem we have that $\|\widehat{\mathbf{S}}_{K_1^C K_2^C}\|_\infty < 1$. This concludes the proof of the theorem.

8.6 Proof of Theorem 6

Without loss of generality assume $\sigma = 1$. From the proof of Theorem 5, it is sufficient to show that $\widehat{\mathbf{S}}_{K_1^C K_2^C}$ cannot be chosen so that $\widehat{\mathbf{K}}_{(K_1^C K_2^C)(K_1^C K_2^C)} \succeq \mathbf{0}$. This is equivalent to showing that

$$\min_{\|\mathbf{S}_{K_1^C K_2^C}\|_\infty \leq 1} \max_{\|\mathbf{x}\|_2 = 1} \mathbf{x}' \left[\begin{pmatrix} \mathbf{0} & \mathbf{A}_{K_1^C K_2^C} \\ \mathbf{A}'_{K_1^C K_2^C} & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \mathbf{S}_{K_1^C K_2^C} \\ \mathbf{S}'_{K_1^C K_2^C} & \mathbf{0} \end{pmatrix} \right] \mathbf{x} > \widehat{\alpha} \quad (20)$$

with probability tending to 1. The left hand side of Eq. (20) is lower bounded by

$$\frac{2\|\mathbf{\Delta}_{K_1^C K_2^C} + \lambda \mathbf{S}_{K_1^C K_2^C}\|_F}{\min(\sqrt{n_1 - k_1}, \sqrt{n_2 - k_2})}.$$

Entries of $\mathbf{A}_{K_1^C K_2^C}$ are soft-thresholded towards zero by $\mathbf{S}_{K_1^C K_2^C}$ to minimize the Frobenious norm. Using (25),

$$\mathbb{P}[|\mathcal{N}(0,1)| > 2\lambda] \geq \tfrac{4\lambda}{\sqrt{2\pi}(4\lambda^2 + 1)} \exp(-2\lambda^2) =: c_\lambda.$$

Using the assumption that $\lambda = \sqrt{c \log \max(n_1 - k_1, n_2 - k_2)}$, we get that $c_\lambda = (\max(n_1 - k_1, n_2 - k_2))^{-2c} L_n$, where $L_n = \mathcal{O}(\text{polylog}(\max(n_1 - k_1, n_2 - k_2)))$.

Let $Z \sim \text{Bin}(N, c_\lambda)$ with $N = (n_2 - k_2)(n_1 - k_1)$. From Lemma 18, $Z > Nc_\lambda/2$ with probability $1 - 2\exp(-Nc_\lambda/8)$. Conditioning on the event $\{Z > Nc_\lambda/2\}$, the left hand side of (20) is lower bounded by

$$\frac{2\lambda\sqrt{2Nc_\lambda}}{\min(\sqrt{n_2 - k_2}, \sqrt{n_1 - k_1})} = 2\lambda\sqrt{2c_\lambda} \max(\sqrt{n_1 - k_1}, \sqrt{n_2 - k_2}).$$

Plugging in the expression for c_λ found above, we see that the left hand side of (20) is lower bounded by $(\max(n_1 - k_1, n_2 - k_2))^{1/2 - c} L_n$.

Lemma 16 provides an upper bound for the right hand side of (20) of the form $\lambda\sqrt{k_1 k_2} + 2(\sqrt{k_1} + \sqrt{k_2})$ with probability $1 - 2\exp(-(\sqrt{k_1} + \sqrt{k_2})^2/2)$. We can conclude that (20) holds with probability tending to one, since

$$(\max(n_1 - k_1, n_2 - k_2))^{1/2 - c} L_n \geq \sqrt{c k_1 k_2 \log \max(n_1 - k_1, n_2 - k_2)} + 2(\sqrt{k_1} + \sqrt{k_2})$$

for sufficiently large n_1 and n_2 as $k_1 = o(n^{1/2 - c})$ and $k_2 = o(n^{1/2 - c})$ under assumptions.

The theorem follows since $\beta = 2\lambda\sqrt{k_1 k_2}$. The constant c can be chosen so that $c < 1/2$.

9 Appendix: Identifying Large Biclusters Without Normality Assumption

We now consider a computationally feasible nonparametric procedure for biclustering that makes minimal assumptions on the distribution of the noise and on the form of the signal. When the clusters are large in a sense specified by the theorem below, the procedure recovers the true bicluster with large probabiliy.

Let F be any distribution with median zero and positive, continuous density. As before, we let $\mathbf{\Delta}$ be a $n_1 \times n_2$ error matrix filled with iid draws from F. We now assume that

$$\mathbf{A} = \mathbf{B} + \mathbf{\Delta}$$

14

where $\mathbf{B} = \{B_{ij}\}_{i \in [n_1], j \in [n_2]}$ is such that $B_{ij} = 0$ for $(i, j) \in K_1 \times K_2$ and

$$\beta \equiv \min_{i \in K_1, j \in K_2} B_{ij} > 0.$$

Let C_j denote the number of positive entries in the jth column of A and let R_i denote the number of positive entries in the tth row of A. Define $\Psi(A)$ to consist of all rows such that $R_i > r \equiv (n_2/2) + \sqrt{n_2 \log n_2}$ and all columns such that $C_j > c \equiv (n_1/2) + \sqrt{n_1 \log n_1}$.

Let $Z \sim F$ and define $\pi = \mathbb{P}(Z + \beta > 0) = 1 - F(\beta)$. Finally, we measure the signal strength by the quantities

$$\psi_1 = k_1 \left[\frac{1}{2} - F(-\beta) \right], \quad \psi_2 = k_2 \left[\frac{1}{2} - F(-\beta) \right].$$

Theorem 8. *Suppose that the following conditions hold:*

$$\psi_1 > \sqrt{4 \log(k_2 n_1)} \tag{21}$$

$$\psi_2 > \sqrt{4 \log(k_1 n_2)}$$

$$\psi_1 \geq \sqrt{n_1 \log n_1} \tag{22}$$

$$\psi_2 \geq \sqrt{n_2 \log n_2}.$$

Then

$$\mathbb{P}(\Psi(A) \neq (K_1, K_2)) \leq 4 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Proof. Consider a null column that does not intersect the cluster. Then $C_j \sim \text{Binomial}(n_1, 1/2)$. By Hoeffding's inequality, $\mathbb{P}(C_j > c) \leq 1/n_1^2$. Similarly for a null row, $\mathbb{P}(R_j > r) \leq 1/n_2^2$. By the union bound, the probability of including any null row or column is at most $n_1/n_1^2 + n_2/n_2^2 = (1/n_1) + (1/n_2)$.

Now consider a non-null column. For simplicity assume that all nonzero β_{ij} are equal to the minimum value β. The extension to the general case is straightforward. Then $C_j = U + V$ where $U \sim \text{Binomial}(n_1 - k_1, 1/2)$ and $V \sim \text{Binomial}(k_1, \pi)$ where $\pi = \mathbb{P}(Z + \beta > 0) = 1 - F(-\beta)$. Here, $Z \sim F$. The probability of excluding column j is $\mathbb{P}(U + V < c)$. Now $U + V$ is the sum of independent but not identically distributed Bernoulli random variables. Applying Hoeffding's inequality for non identically distributed variables we have $\mathbb{P}(U + V < c) \leq e^{-2(\mu - c)^2/n_1}$ where

$$\mu = \mathbb{E}(U + V) = \frac{n_1 - k_1}{2} + k_1 \pi.$$

Substituting for μ and c and using the fact that $\pi - 1/2 = 1/2 - F(-\beta)$,

$$
\begin{aligned}
\mathbb{P}(U + V < c) &\leq e^{-2(\mu - c)^2/n_1} \\
&= \exp\left(\frac{k_1}{\sqrt{n_1}}(\pi - 1/2) - \frac{1}{2}\sqrt{\log n_1} \right)^2 \\
&\leq \exp\left(-\frac{k_1^2 (\pi - 1/2)^2}{4 n_1} \right)
\end{aligned}
$$

where we used (22). By (21), the last quantity is less than $1/(k_2 n_1)$. Taking the union bound over all the k_2 columns in the cluster, the probability of missing a relevant column is at most $1/n_1$. A similar bound applies to the rows. $\qquad \square$

10 Appendix: Technical proofs

10.1 Proof of Lemma 7

It follows directly from Weyl's theorem (e.g. [25]) that

$$|\alpha - \widehat{\alpha}| \leq \sigma_1(\mathbf{\Delta}_{K_1 K_2}). \tag{23}$$

Denote $\widehat{\mathbf{u}}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}$ the singular vectors of $\mathbf{M}$ associated with $\widehat{\alpha}$, that is,

$$
\begin{aligned}
\mathbf{M}\widehat{\mathbf{v}}_{K_2} &= \widehat{\alpha}\widehat{\mathbf{u}}_{K_1}, \quad \text{and} \\
\mathbf{M}'\widehat{\mathbf{u}}_{K_1} &= \widehat{\alpha}\widehat{\mathbf{v}}_{K_2}.
\end{aligned}
\tag{24}
$$

Let $\mathbf{u}_{K_1}^{\perp} \in \{\mathbf{a} \in \mathbb{R}^{k_1} : \mathbf{a} \perp \mathbf{u}_{K_1}, , \|\mathbf{a}\| = 1\}$ and $\mathbf{v}_{K_2}^{\perp} \in \{\mathbf{a} \in \mathbb{R}^{k_2} : \mathbf{a} \perp \mathbf{v}_{K_2}, \|\mathbf{a}\| = 1\}$. With this we write $\widehat{\mathbf{v}}_{K_2} = c_1^v \mathbf{v}_{K_2} + c_0^v \mathbf{v}_{K_2}^{\perp}$ and $\widehat{\mathbf{u}}_{K_1} = c_1^u \mathbf{u}_{K_1} + c_0^u \mathbf{u}_{K_1}^{\perp}$ where $(c_1^v)^2 + (c_0^v)^2 = 1$ and $(c_1^u)^2 + (c_0^u)^2 = 1$. Lemma 9 gives a lower bound on c_1^u and c_1^v and is proven below.

From (24) we have

$$
\alpha c_1^v \mathbf{u}_{K_1} + \boldsymbol{\Delta}_{K_1 K_2}\widehat{\mathbf{v}}_{K_2} = \widehat{\alpha}\widehat{\mathbf{u}}_{K_1}
$$

which further decomposes into

$$
\alpha c_1^v \mathbf{u}_{K_1} + \boldsymbol{\Delta}_{K_1 K_2}(c_1^v \mathbf{v}_{K_2} + c_0^v \mathbf{v}_{K_2}^{\perp}) = \widehat{\alpha}(\widehat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}) + \widehat{\alpha}\mathbf{u}_{K_1}.
$$

Using Taylor series expansion $\widehat{\alpha}^{-1} \lesssim \alpha^{-1} + \sigma_1(\boldsymbol{\Delta}_{K_1 K_2})\alpha^{-2}$. Now

$$
\begin{aligned}
\|\widehat{\mathbf{u}}_{K_1} &- \mathbf{u}_{K_1}\|_{\infty} \\
&\leq |\widehat{\alpha}^{-1}\alpha c_1^v - 1|\|\mathbf{u}_{K_1}\|_{\infty} + \widehat{\alpha}^{-1}|c_1^v|\|\boldsymbol{\Delta}_{K_1 K_2}\mathbf{v}_{K_2}\|_{\infty} + \widehat{\alpha}^{-1}|c_0^v|\|\boldsymbol{\Delta}_{K_1 K_2}\mathbf{v}_{K_2}^{\perp}\|_{\infty} + o(1) \\
&\leq 2\alpha^{-1}\sigma_1(\boldsymbol{\Delta}_{K_1 K_2})\|\mathbf{u}_{K_1}\|_{\infty} + \alpha^{-1}\|\boldsymbol{\Delta}_{K_1 K_2}\mathbf{v}_{K_2}\|_{\infty} + 2\alpha^{-2}\sigma_1(\boldsymbol{\Delta}_{K_1 K_2})\|\boldsymbol{\Delta}_{K_1 K_2}\|_{\infty,2} + o(1)
\end{aligned}
$$

using (23) and Lemma 9. The three terms in the display above can be bounded using Lemma 16, Lemma 13 and Lemma 14. Then

$$
\|\widehat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_{\infty} = \alpha^{-1}\mathcal{O}\left(\sqrt{k_1}\|\mathbf{u}_{K_1}\|_{\infty} + \sqrt{\log k_1} + \alpha^{-1}k_2\right) = \alpha^{-1}\mathcal{O}(\sqrt{\log k_1})
$$

with probability $1 - \mathcal{O}(k_1^{-1})$. A similar calculation gives a bound on $\|\widehat{\mathbf{v}}_{K_2} - \mathbf{v}_{K_2}\|_{\infty}$. This completes the proof of Lemma 7.

The following Lemma establishes a lower bound on $\widehat{\mathbf{u}}_{K_1}'\mathbf{u}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}'\mathbf{v}_{K_2}$ under our sign convection.

Lemma 9. *We have that* $c_1^u \geq 1 - 2\alpha^{-1}\sigma_1(\boldsymbol{\Delta}_{K_1 K_2})$ *and* $c_1^v \geq 1 - 2\alpha^{-1}\sigma_1(\boldsymbol{\Delta}_{K_1 K_2})$.

Proof of Lemma 9. From (24) we have

$$
\alpha\widehat{\mathbf{u}}_{K_1}'\mathbf{u}_{K_1}\mathbf{v}_{K_2}'\widehat{\mathbf{v}}_{K_2} + \widehat{\mathbf{u}}_{K_1}'\boldsymbol{\Delta}_{K_1 K_2}\widehat{\mathbf{v}}_{K_2} = \widehat{\alpha}.
$$

Using the triangle inequality

$$
\begin{aligned}
|\alpha - \alpha\widehat{\mathbf{u}}_{K_1}'\mathbf{u}_{K_1}\mathbf{v}_{K_2}'\widehat{\mathbf{v}}_{K_2}| &\leq |\alpha - \widehat{\alpha}| + |\widehat{\alpha} - \alpha\widehat{\mathbf{u}}_{K_1}'\mathbf{u}_{K_1}\mathbf{v}_{K_2}'\widehat{\mathbf{v}}_{K_2}| \\
&\leq 2\sigma_1(\boldsymbol{\Delta}_{K_1 K_2}),
\end{aligned}
$$

since $|\widehat{\mathbf{u}}_{K_1}'\boldsymbol{\Delta}_{K_1 K_2}\widehat{\mathbf{v}}_{K_2}| \leq \sigma_1(\boldsymbol{\Delta}_{K_1 K_2})$. Under our sign convention, this implies that

$$
1 - \widehat{\mathbf{u}}_{K_1}'\mathbf{u}_{K_1}\mathbf{v}_{K_2}'\widehat{\mathbf{v}}_{K_2} \leq 2\alpha^{-1}\sigma_1(\boldsymbol{\Delta}_{K_1 K_2}).
$$

We conclude that

$$
\begin{aligned}
\widehat{\mathbf{u}}_{K_1}'\mathbf{u}_{K_1} &\geq 1 - 2\alpha^{-1}\sigma_1(\boldsymbol{\Delta}_{K_1 K_2}), \quad \text{and} \\
\widehat{\mathbf{v}}_{K_1}'\mathbf{v}_{K_1} &\geq 1 - 2\alpha^{-1}\sigma_1(\boldsymbol{\Delta}_{K_1 K_2}).
\end{aligned}
$$

$\square$

11 Appendix: Collection of concentration results

In this section, we collect useful results on tail bounds of various random quantities used throughout the paper. We start by stating a lower and upper bound on the survival function of the standard normal random variable. Let $Z \sim \mathcal{N}(0,1)$ be a standard normal random variable. Then for $t > 0$

$$
\frac{1}{\sqrt{2\pi}}\frac{t}{t^2 + 1}\exp(-t^2/2) \leq \mathbb{P}(Z > t) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}\exp(-t^2/2).
\tag{25}
$$

We will use the above inequality to bound some quantities involving norms of random matrices with independent standard normal entries. We provide a few more definitions.

Definition 10. *Let ϵ be a positive number. A set X is an ϵ-net of a set Y if for any $y \in Y$, there exists $x \in X$ such that $||y - x|| \leq \epsilon$.*

The following result is the standard ϵ-net argument.

Lemma 11. *Let $\mathcal{N} \subset \mathcal{S}^{n_2 - 1}$ be an ϵ-net $\mathcal{N}$ of $\mathcal{S}^{n_2 - 1}$ and let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a linear map. Then there is a vector $\mathbf{y} \in \mathcal{N}$ such that*

$$||\mathbf{A}\mathbf{y}|| \geq (1 - \epsilon) \max_{\mathbf{x} \in \mathcal{S}^{n_2 - 1}} ||\mathbf{A}\mathbf{x}||.$$

The minimum size of the ϵ-net is well-known.

Lemma 12. *There is an ϵ-net of a unit sphere in d dimensions of size at most $(\frac{3}{\epsilon})^d$.*

Lemma 13. *Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. Then for any fixed $x \in \mathcal{S}^{n_2 - 1}$,*

$$\mathbb{P}[||\mathbf{A}\mathbf{x}||_\infty \geq t] \leq \frac{2n_1}{\sqrt{2\pi}t} \exp(-t^2/2).$$

Proof of Lemma 13. Observe that $\mathbf{A}\mathbf{x} \sim N(0, \mathbf{I}_{n_1})$. The result follows from an application of a standard Gaussian tail bound and the union bound. $\qquad\square$

The following two results bound operator norms $\|\mathbf{A}\|_{\infty,2}$ and $\|\mathbf{A}\|_{\infty,\infty}$.

Lemma 14. *Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. Fix $\delta > 0$. Then*

$$\|\mathbf{A}\|_{\infty,2} \leq \sqrt{8\left(\log n_1 + n_2 \log 6 + \log \frac{2}{\sqrt{2\pi}\delta}\right)} =: K_{\delta, n_1, n_2} \tag{26}$$

with probability $1 - \delta/K_{\delta, n_1, n_2}$.

Proof of Lemma 14. By definition, we have that

$$\|\mathbf{A}\|_{\infty,2} = \max_{||\mathbf{x}||_2 \leq 1} ||\mathbf{A}\mathbf{x}||_\infty.$$

Let $\mathcal{N} \subset \mathcal{S}^{n_2}$ be an ϵ-net of $\mathcal{S}^{n_2 - 1}$. Using Lemma 11 we have that

$$\mathbb{P}[\|\mathbf{A}\|_{\infty,2} \geq t] \leq \mathbb{P}[(1 - \epsilon)^{-1} \max_{\mathbf{y} \in \mathcal{N}} ||\mathbf{A}\mathbf{y}||_\infty \geq t].$$

Setting $\epsilon = \frac{1}{2}$, applying Lemma 12, Lemma 13 and using the union bound, we have that

$$\mathbb{P}[\|\mathbf{A}\|_{\infty,2} \geq t] \leq \frac{2n_1}{\sqrt{2\pi}t} 6^{n_2} \exp(-t^2/8).$$

We can conclude the proof by setting $t = K_{\delta, n_1, n_2}$. $\qquad\square$

Lemma 15. *Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. Fix $\delta > 0$. Then there exists a sufficiently large constant C such that*

$$\|\mathbf{A}\|_{\infty,\infty} \leq \sqrt{8\left(n_2 \log n_1 + n_2^2 \log 6 + n_2 \log \frac{2}{\sqrt{2\pi}\delta}\right)} =: \sqrt{n_2} K_{\delta, n_1, n_2} \tag{27}$$

with probability $1 - \delta/K_{\delta, n_1, n_2}$ where K_{δ, n_1, n_2} is defined in (26).

Proof of Lemma 15. For any $\mathbf{x} \in \mathbb{R}^{n_2}$, $||\mathbf{x}||_2 \leq \sqrt{k}||\mathbf{x}||_\infty$. Now

$$\|\mathbf{A}\|_{\infty,\infty} = \max_{||\mathbf{x}||_\infty \leq 1} ||\mathbf{A}\mathbf{x}||_\infty \leq \max_{||\mathbf{x}||_2 \leq \sqrt{n_2}} ||\mathbf{A}\mathbf{x}||_\infty = \sqrt{n_2}\|\mathbf{A}\|_{\infty,2}.$$

The result follows from Lemma 14. $\qquad\square$

Lemma 16 ([12]). *Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. We have that*

$$\mathbb{P}[\sigma_1(\mathbf{A}) \geq \sqrt{n_1} + \sqrt{n_2} + t] \leq 2\exp(-t^2/2). \tag{28}$$

Lemma 17. *If $z_k \sim \mathrm{Bin}(k, \pi_k)$, then for all $k \geq 1$ and all $\pi_k \in (0, 1)$ it holds that*

$$\mathbb{P}[z_k = 0] \leq \exp(-k\pi_k).$$

Proof. $\mathbb{P}[z_k = 0] = (1-\pi_k)^k = \exp(-k\log(\frac{1}{1-\pi_k})) = \exp(-k(\pi_k+\mathcal{O}(\pi_k^2))) \leq \exp(-k\pi_k).$ □

Lemma 18. *If $z_k \sim \mathrm{Bin}(k, \pi_k)$, then*

$$\mathbb{P}[z_k \leq k\pi_k - t] \leq \exp(-t^2/(2k\pi_k))$$

and

$$\mathbb{P}[z_k \geq k\pi_k + t] \leq \exp(-t^2/(2(k\pi_k + t/3))).$$

12 Appendix: Convex analysis

The following results are standard. We use them to derive the KKT condition for the optimization problem in (11).

Definition 19. *Let $\mathcal{C}$ be a convex set. The function $\delta(x|\mathcal{C})$ defined as*

$$\delta(x|\mathcal{C}) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ +\infty & \text{if } x \notin \mathcal{C} \end{cases}$$

is called the indicator function of the convex set $\mathcal{C}$.

Definition 20. *Let $\partial\delta(x|\mathcal{C})$ denote the normal cone to $\mathcal{C}$ at x defined as*

$$\partial\delta(a|\mathcal{C}) = \{y : \langle x - a, y\rangle \leq 0, \ \forall x \in \mathcal{C}\}.$$

The normal cone be equivalently defined as

$$\partial\delta(a|\mathcal{C}) = \{y : \sup_{x \in \mathcal{C}}\langle x, y\rangle = \langle a, y\rangle\}.$$

If a is interior to $\mathcal{C}$ then $\partial\delta(a|\mathcal{C}) = \{0\}$, and if a is exterior to $\mathcal{C}$ then $\partial\delta(a|\mathcal{C}) = \emptyset$

Let $\mathcal{S}_+^n$ be the cone of positive semi-definite symmetric matrices in $\mathbb{R}^{n \times n}$.

Theorem 21 ([13]). *The normal cone to $\mathcal{S}_+^n$ is defined as*

$$\partial\delta(\mathbf{A}|\mathcal{S}_+^n) = \begin{cases} \emptyset & \text{if } \mathbf{A} \notin \mathcal{S}_+^n \\ \{\mathbf{B} \ : \ -\mathbf{B} \in \mathcal{S}_+^n, \ \mathrm{tr}\,\mathbf{AB} = 0\} & \text{if } \mathbf{A} \in \mathcal{S}_+^n. \end{cases} \tag{29}$$

Alternatively for $\mathbf{A} \in \mathcal{S}_+^n$, equation (29) becomes

$$\partial\delta(\mathbf{A}|\mathcal{S}_+^n) = \{\mathbf{B} = -\mathbf{Z}\Lambda\mathbf{Z}' \ : \ \Lambda \in \mathcal{S}_+^n\} \tag{30}$$

where columns of $\mathbf{Z}$ form orthonormal basis for the null space of $\mathbf{A}$.

Theorem 22 ([23], Chapter 5). *If $\widehat{\mathbf{A}}$ solves the problem*

$$\begin{array}{ll} \min & f(\mathbf{A}) \\ \text{subject to} & \mathbf{A} \in \mathcal{S}_+^n, \quad g(\mathbf{A}) \leq 0, \end{array}$$

then $\widehat{\mathbf{A}}$ is feasible and there exist matrices $\widehat{\mathbf{G}} \in \partial f(\widehat{\mathbf{A}})$, $\widehat{\mathbf{B}} \in \partial\delta(\widehat{\mathbf{A}}|\mathcal{S}_+^n)$, $\mathbf{C} \in \partial g(\widehat{\mathbf{A}})$ and a multiplier $\widehat{\pi} \geq 0$, $\widehat{\pi}g(\widehat{\mathbf{A}}) = 0$ such that

$$\widehat{\mathbf{G}} + \widehat{\mathbf{B}} + \widehat{\pi}\widehat{\mathbf{C}} = 0.$$

13 Appendix: Nuclear norm and ℓ_1 norm penalty

Under the model (1), the problem of biclustering can be thought of recovering a matrix that is both low rank and sparse. As pointed out by a reviewer, from this point of view a natural combination of the nuclear norm and the ℓ_1 norm leads to the following optimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}} \frac{1}{2}||\mathbf{A}-\mathbf{X}||_F^2 + \lambda_1||\mathbf{X}||_* + \lambda_2\mathbf{1}'|\mathbf{X}|\mathbf{1}. \tag{31}$$

The norm $||\mathbf{X}||_*$ is the nuclear norm defined as the sum of the singular values of $\mathbf{X}$, that is, if $\mathbf{X}=\mathbf{UDV}'$ is the singular value decomposition of $\mathbf{X}$, then $||\mathbf{X}||=\sum_i D_{ii}$. The tuning parameter λ_1 control the rank of the solution and λ_2 controls the sparsity of the solution. Compared to the optimization procedure in (11), there is an additional tuning parameter that needs to be selected in practice. Combination of the nuclear norm and the ℓ_1 norm was shown useful in robust PCA [10]. For the problem of biclustering, the formulation in (31) does not lead to improvement over (11) as we show below.

We analyze the problem (31) in a similar way to the proof of Theorem 5. That is, we construct a rank one solution $\widehat{\mathbf{X}}$ that is a global solution of the objective (31). The following Lemma gives a subgradient of the nuclear norm used in stating the first order conditions for a global optimum.

Lemma 23. *If* $\mathbf{X}=\mathbf{UDV}'$ *is the singular value decomposition of* $\mathbf{X}$ *then the subdifferential of* $||\cdot||_*$ *is equal to*

$$\partial||\mathbf{X}||_* = \{\mathbf{UV}'+\mathbf{Z} \ : \ \sigma_1(\mathbf{Z})\le 1, \ \mathbf{U}'\mathbf{Z}=0 \text{ and } \mathbf{ZV}=0\}. \tag{32}$$

Now, the first order condition for a global optimum of (31) is

$$\widehat{\mathbf{X}} - \mathbf{A} + \lambda_1\widehat{\mathbf{K}} + \lambda_2\widehat{\mathbf{S}} = \mathbf{0} \tag{33}$$

where $\widehat{\mathbf{S}}\in\partial||\widehat{\mathbf{X}}||_1$ and $\widehat{\mathbf{K}}\in\partial||\widehat{\mathbf{X}}||_*$.

Suppose that the matrix $\widehat{\mathbf{X}}$ is rank one and that the sparsity pattern of $\widehat{\mathbf{X}}$ correctly recovers K_1 and K_2. Denote $\widehat{\mathbf{X}}=\widehat{\alpha}\widehat{\mathbf{u}}\widehat{\mathbf{v}}'$. Then we have that $\widehat{\mathbf{S}}_{K_1K_2}=\mathbf{s}_{\widehat{\mathbf{u}}}\mathbf{s}'_{\widehat{\mathbf{v}}}$ where $\mathbf{s}_{\widehat{\mathbf{u}}}=\text{sign}(\widehat{\mathbf{u}}_{K_1})$ and $\mathbf{s}_{\widehat{\mathbf{v}}}=\text{sign}(\widehat{\mathbf{v}}_{K_2})$. Furthermore, from Lemma 23, we know that $\widehat{\mathbf{K}}=\widehat{\mathbf{u}}\widehat{\mathbf{v}}'+\widehat{\mathbf{Z}}$ with $\sigma_1(\mathbf{Z})\le 1$, $\widehat{\mathbf{u}}'\mathbf{Z}=0$ and $\mathbf{Z}\widehat{\mathbf{v}}=0$.

Observe that the problem (31) can be rewritten as

$$\max_{\mathbf{X}\in\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}} \text{tr }\mathbf{A}'\mathbf{X} - \frac{1}{2}\text{tr }\mathbf{X}'\mathbf{X} - \lambda_1||\mathbf{X}||_* - \lambda_2\mathbf{1}'|\mathbf{X}|\mathbf{1}.$$

Under the assumption that $\widehat{\mathbf{X}}=\widehat{\alpha}\widehat{\mathbf{u}}\widehat{\mathbf{v}}'$ with $\widehat{\mathbf{u}}=(\widehat{\mathbf{u}}'_{K_1},\mathbf{0}')'$ and $\widehat{\mathbf{v}}=(\widehat{\mathbf{v}}'_{K_2},\mathbf{0}')'$, the above equation becomes

$$\max_{\widehat{\alpha},\widehat{\mathbf{u}}_{K_1},\widehat{\mathbf{v}}_{K_2}} \widehat{\alpha}\widehat{\mathbf{u}}'_{K_1}\mathbf{A}_{K_1K_2}\widehat{\mathbf{v}}'_{K_2}-\widehat{\alpha}^2-\frac{1}{2}\lambda_1\widehat{\alpha}-\lambda_2\widehat{\alpha}\widehat{\mathbf{u}}'_{K_1}\mathbf{s}_{\widehat{\mathbf{u}}}\mathbf{s}'_{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}_{K_2} \quad \text{subject to } ||\widehat{\mathbf{u}}_{K_1}||_2=1, \ ||\widehat{\mathbf{v}}_{K_2}||_2=1. \tag{34}$$

The objective (31) is strongly convex, which implies that $\widehat{\alpha}$, $\widehat{\mathbf{u}}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}$ are unique if the global solution is of rank one. This in turn implies that $\widehat{\mathbf{u}}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}$ are left and right singular vectors of $\mathbf{A}_{K_1K_2}-\lambda_2\mathbf{s}_{\widehat{\mathbf{u}}}\mathbf{s}'_{\widehat{\mathbf{v}}}$. Setting $\lambda_2=\frac{\beta}{2\sqrt{k_1k_2}}$ and $\alpha=\beta/2$, we observe that the results of Lemma 7 hold here. That is, under the conditions of Theorem 5, it holds that

$$||\widehat{\mathbf{u}}_{K_1}-\mathbf{u}_{K_1}||_\infty = \mathcal{O}\left(\sqrt{\frac{\log k_1}{k_1k_2\log(n_1-k_2)(n_2-k_2)}}\right)$$

and

$$||\widehat{\mathbf{v}}_{K_2}-\mathbf{v}_{K_2}||_\infty = \mathcal{O}\left(\sqrt{\frac{\log k_2}{k_1k_2\log(n_1-k_2)(n_2-k_2)}}\right)$$

with probability $1-\mathcal{O}(k_1^{-1})$. With $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ fixed, the problem (34) can be explicitly solved for $\widehat{\alpha}$,

$$\widehat{\alpha} = \sigma_1(\alpha\mathbf{u}_{K_1}\mathbf{v}'_{K_2}+\mathbf{\Delta}_{K_1K_2}) - \lambda_1, \tag{35}$$

which gives us a constraint on the signal strength α and the tuning parameter λ_1.

So far, we have constructed $\widehat{\mathbf{X}}_{K_1 K_2}$ and $\widehat{\mathbf{S}}_{K_1 K_2}$. We need to verify that there is a matrix $\widehat{\mathbf{Z}}$ that satisfies (32) by plugging back $\widehat{\mathbf{X}}_{K_1 K_2}$ and $\widehat{\mathbf{S}}_{K_1 K_2}$ into (33). We will construct

$$\widehat{\mathbf{Z}} = \begin{pmatrix} \widehat{\mathbf{Z}}_{K_1 K_2} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{Z}}_{K_1^C K_2^C} \end{pmatrix}. \tag{36}$$

From (33), we observe that

$$(\widehat{\alpha} + \lambda_1)\widehat{\mathbf{u}}_{K_1}\widehat{\mathbf{v}}_{K_2}' - \alpha \mathbf{u}_{K_1}\mathbf{v}_{K_2}' - \mathbf{\Delta}_{K_1 K_2} = \lambda_1 \widehat{\mathbf{Z}}_{K_1 K_2}.$$

It follows that we need $\lambda_1 = \Omega(\sqrt{k_1} + \sqrt{k_2})$ to ensure that $\sigma_1(\widehat{\mathbf{Z}}_{K_1 K_2}) \leq 1$.

We have already seen in the proof of Theorem 5 that $\widehat{\mathbf{S}}_{K_1^C K_2} = \lambda_2^{-1} \mathbf{\Delta}_{K_1^C K_2}$, $\widehat{\mathbf{S}}_{K_1 K_2^C} = \lambda_2^{-1} \mathbf{\Delta}_{K_1 K_2^C}$ and $\widehat{\mathbf{S}}_{K_1^C K_2^C} = \lambda_2^{-1} \mathbf{\Delta}_{K_1^C K_2^C}$ are valid blocks of a subdifferential of the ℓ_1 norm. Plugging back into (33), it follows that $\widehat{\mathbf{Z}}_{K_1^C K_2^C} = \mathbf{0}$.

We can conclude that under the conditions of Theorem 5 on the size of the bicluster and the signal strength β with $\lambda_1 = \mathcal{O}(\sqrt{k_1} + \sqrt{k_2})$ and $\lambda_2 = \frac{\beta}{2\sqrt{k_1 k_2}}$, the solution $\widehat{\mathbf{X}}$ of (31) is of rank one and correctly recovers (K_1, K_2).

We can also show a similar result to Theorem 6, which establishes that the signal strength β cannot be much smaller than the one given in Theorem 5. From (33) follows that

$$\sigma_1(\mathbf{\Delta}_{K_1^C K_2^C} - \lambda_2 \widehat{\mathbf{S}}_{K_1^C K_2^C}) \leq \lambda_1$$

is necessary for $\widehat{\mathbf{X}}$ to be of rank one and to correctly recover (K_1, K_2). From (35), we have that $\lambda < \sigma_1(\alpha \mathbf{u}_{K_1}\mathbf{v}_{K_2}' + \mathbf{\Delta}_{K_1 K_2})$ for a solution to be of rank 1. Since $\sigma_1(\alpha \mathbf{u}_{K_1}\mathbf{v}_{K_2}' + \mathbf{\Delta}_{K_1 K_2}) < \alpha + 2(\sqrt{k_1} + \sqrt{k_2})$ with high probability, we have that $\lambda < \alpha + 2(\sqrt{k_1} + \sqrt{k_2})$. However, it was shown in the proof of Theorem 6 that

$$\min_{||\mathbf{S}_{K_1^C K_2^C}||_\infty \leq 1} \max_{||\mathbf{x}||_2 = 1} \mathbf{x}' \left[\begin{pmatrix} \mathbf{0} & \mathbf{A}_{K_1^C K_2^C} \\ \mathbf{A}'_{K_1^C K_2^C} & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \mathbf{S}_{K_1^C K_2^C} \\ \mathbf{S}'_{K_1^C K_2^C} & \mathbf{0} \end{pmatrix} \right] \mathbf{x}$$
$$> \alpha + 2(\sqrt{k_1} + \sqrt{k_2}) \tag{37}$$

with probability tending to 1.