Statistical and computational tradeoffs in biclustering

Sivaraman Balakrishnan[†] sbalakri@cs.cmu.edu
Mladen Kolar[†] mladenk@cs.cmu.edu
Alessandro Rinaldo[††] arinaldo@stat.cmu.edu
Aarti Singh[†] aarti@cs.cmu.edu
Larry Wasserman[††] larry@stat.cmu.edu

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

Abstract

We consider the problem of identifying a small sub-matrix of activation in a large noisy matrix. We establish the minimax rate for the problem by showing tight (up to constants) upper and lower bounds on the signal strength needed to identify the sub-matrix. We consider several natural computationally tractable procedures and show that under most parameter scalings they are unable to identify the sub-matrix at the minimax signal strength. While we are unable to directly establish the computational hardness of the problem at the minimax signal strength we discuss connections to some known NP-hard problems and their approximation algorithms.

1 Introduction

We are given an $(n \times n)$ matrix M, which is a sparse $(k \times k)$ matrix of activation, of strength at least $\mu_{\min}$, drowned in a large noisy matrix. More formally let,

$$\Theta(\mu_{\min}, n, k) := \{(\mu, K_1, K_2) \ : \ \mu \geq \mu_{\min}, |K_1| = k, K_1 \subset [n], |K_2| = k, K_2 \subset [n]\} \tag{1}$$

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

$$\prod_{ij} \mathcal{N}(m_{ij}; \mu \mathbb{1}\{i \in K_1, j \in K_2\}, \sigma^2), \tag{2}$$

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 assume that μ is positive, and that k and σ are known. Without loss of generality σ can be taken to be 1 (by appropriate rescaling).

Given the matrix M, the problem is to identify the rows and columns which together constitute the sub-matrix of activation. We are interested in identifying tuples of $(n, k, \mu_{\min})$ for which this problem is both statistically feasible and computationally tractable.

2 Related work

In many real world applications it is necessary to identify small sub-matrices in a large noisy matrix. A prototypical example is bi-clustering, which is the problem of identifying a (typically) sparse set of relevant columns and rows in a large, noisy data matrix. These columns and rows define a sub-matrix that needs to be identified. Due to its practical importance and difficulty bi-clustering has attracted considerable attention [5, 9, 10, 12]. 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. [11] provide some statistical backing for these exhaustive search procedures.

The problem setup introduced in [8] is also related to 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 sub-matrix however imposes structure on which elements are elevated concurrently. Several other recent papers [1, 2, 3] have investigated the structured normal means setting for different ordered domains.

3 Known results

We have addressed the bi-clustering problem in recent work [8]. We summarize the relevant results of that paper here. Define an *estimator* Ψ as any function that takes M as input and outputs its estimate of the coordinates of the matrix which are active (these may or may not form a sub-matrix). We analyze the probability that these coordinates are *exactly* the true active sub-matrix. This is a $0/1$ loss and while many of our results can be transferred to other losses (like the Hamming loss) we will focus only on the $0/1$ loss in this paper.

We are interested in a minimax analysis of the bi-clustering problem. To compute the minimax probability of error we consider the supremum over all matrices M drawn according to $\theta \in \Theta$ and the infimum over all estimators Ψ. First, we have an information theoretic lower bound on the signal strength needed for *any* procedure to succeed.

Theorem 1. *There exists a constant c (small) such that if* $\mu_{\min} \leq c\sqrt{\frac{\log(n-k)}{k}}$ *then for* any *estimator* $\Psi(M)$ *of the bicluster, the* minimax *probability of error remains bounded away from 0.*

To establish this as the minimax rate we need to establish a complementing (upto constants) upper bound. This can be achieved by a procedure that looks over all possible sub-matrices of size k and outputs the one with the largest sum. For two subsets, $\tilde{K}_1 \subset [n]$ and $\tilde{K}_2 \subset [n]$, we define the score $\mathcal{S}(\tilde{K}_1, \tilde{K}_2) := \sum_{i \in \tilde{K}_1} \sum_{j \in \tilde{K}_2} m_{ij}$. Furthermore, define

$$\Psi_{\max}(M) := \underset{(\tilde{K}_1, \tilde{K}_2)}{\operatorname{argmax}} \ \mathcal{S}(\tilde{K}_1, \tilde{K}_2) \quad \text{subject to} \quad |\tilde{K}_1| = k, \ |\tilde{K}_2| = k, \tag{3}$$

for which we have the following result.

Theorem 2. *If the signal strength* $\mu_{\min} \geq 4\sqrt{\frac{\log(n-k)}{k}}$ *then* $\mathbb{P}[\Psi_{\max}(M) \neq (K_1, K_2)] \leq 4[(n-k)^{-1}]$.

This upper bound establishes the minimax signal strength (upto constants). Unfortunately directly solving the combinatorial optimization problem in (3) is intractable. Therefore, we analyze a few computationally tractable procedures.

The simplest procedure is based on element-wise thresholding. The sub-matrix is estimated as

$$\Psi_{\text{thr}}(M, \tau) := \{(i, j) \in [n] \times [n] : m_{ij} \geq \tau\} \tag{4}$$

where $\tau > 0$ is a parameter.

Theorem 3. *Set the threshold* $\tau = \frac{C}{2}\sqrt{\log(n-k)})$ *If*

$$\mu_{\min} \geq C\sqrt{\log(n-k)}$$

then $\mathbb{P}[\Psi_{\text{thr}}(M, \tau) \neq K_1 \times K_2] = o(k^{-2})$ *for C large enough.*

From this result, we observe that the signal strength μ needs to be $O(\sqrt{k})$ larger than the lowest possible signal strength. This is not surprising, since element-wise thresholding is not exploiting the structure of the bicluster, but is assuming that the large elements of the matrix M are positioned randomly. We will refer to the $\mu_{\min}$ from this theorem as the *thresholding* signal strength.

We also consider a procedure based on row and column averaging, that is, we find a sub-matrix defined by the k rows and k columns with the largest average. Denote this estimator $\Psi_{\text{avg}}(M)$. We have the following result.

Theorem 4. *If* $k = \Omega(n^{1/2+\alpha})$*, where* $\alpha \in [0, 1/2]$ *is a constant and,*

$$\mu_{\min} \geq 4\frac{\sqrt{\log(n-k)}}{n^{\alpha}}$$

then $\mathbb{P}[\Psi_{\text{avg}}(M) \neq (K_1, K_2)] \leq [2n^{-1}]$.

Comparing to Theorem 3, we observe that the averaging requires lower signal strength than the element-wise thresholding whenever the bicluster is large, that is, $k = \Omega(\sqrt{n})$. Unless $k = \mathcal{O}(n)$, the procedure does not achieve the lower bound of Theorem 1, however, the procedure is computationally efficient.

Finally, when the submatrix is all constant, the noiseless matrix is low-rank and has sparse singular vectors that pick out the bicluster. We investigate sparse singular value decomposition on the noisy matrix M. Inspired by [7], we investigate the following convex problem:

$$\max_{\mathbf{X} \in \mathbb{R}^{(2n)\times(2n)}} \operatorname{tr} M\mathbf{X}^{21} - \lambda \mathbf{1}'|\mathbf{X}^{21}|\mathbf{1} \quad \text{subject to} \quad \mathbf{X} \succeq \mathbf{0},\ \operatorname{tr}\mathbf{X}^{11} = 1, \operatorname{tr}\mathbf{X}^{22} = 1, \tag{5}$$

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

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

The solution $\mathbf{X}^{21}$ is sparse and recovers the position of the sub-matrix.

Theorem 5. *If*

$$\mu_{\min} \geq 2\sqrt{\log(n-k)} \tag{6}$$

then with $\lambda = \frac{\mu}{2}$ *the recovered submatrix* $(\hat{K}_1, \hat{K}_2) = (K_1, K_2)$ *with probability* $1 - \mathcal{O}(k^{-1})$.

We have further shown in [8] that the result of Theorem 5 cannot be improved. This is somewhat surprising, since the signal strength needed for the success of the procedure is of the same order as for the element-wise thresholding, where from the formulation of the optimization problem it seems that the procedure uses the structure of the problem.

Tradeoffs: The computational and statistical tradeoffs in the above theorems are clear. Below the minimax signal strength threshold the problem is statistically infeasible. Just above this threshold the problem is statistically feasible but appears to be computationally hard for small biclusters. Well above this signal strength (at the thresholding signal strength) the problem is both statistically feasible and computationally tractable. For large biclusters however, the picture is a bit different. As the size of the bicluster grows we are able to approach the minimax threshold with a computationally efficient procedure.

4 New connections and results

In this section we discuss connections between the biclustering problem and certain structured (and random) instances of known NP-hard problems. The combinatorial procedure described above is equivalent to the following ℓ_0 constrained optimization problem.

$$\begin{aligned} \text{maximize}_{u,v} \quad & u^T M v \\ \text{subject to} \quad & u, v \in \{0,1\}^n \\ & ||u||_0 \leq k \\ & ||v||_0 \leq k \end{aligned}$$

The biclustering problem is also a special case of the Quadratic Assignment Problem (QAP), where the objective is given two matrices A and B (typically of weights and distances between facilities and locations) to maximize their Frobenius inner product. In the biclustering problem, the matrix A is just M and B is a $(k \times k)$ matrix of 1s padded appropriately with 0s.

Biclustering is also related to finding the densest-2k subgraph in a bi-partite graph, where we view the matrix M as defining a weighted bi-partite graph with $2n$ nodes and edge weights given by the entries of M. The density of a subgraph is the ratio of the sum of the weights of edges between nodes of the subgraph to the total number of vertices in the subgraph. It is straightforward to show that even at the minimax signal strength the densest 2k-subgraph is the bicluster of interest. The QAP and densest k-subgraph problems are both however NP-hard in general.

The closely related densest subgraph problem is tractable. It can be cast as an ILP [6], whose relaxation can be rounded to get the exact optimal integral solution. It is then interesting to ask at what scalings are the solutions to the densest subgraph and densest 2k-subgraph problems identical. This is also related to the question: what is the size and density of the densest subgraph in a random Gaussian matrix of size $(n \times n)$? We have recently analyzed this question and can show that the solution to densest subgraph problem on the graph induced by M is the true submatrix of activation at signal strengths similar to those at which row/column averaging succeeds with high probability. At the minimax signal strength the bicluster needs

to be of size $O(n)$ and at the thresholding scaling the bicluster needs to be of size at least $O(\sqrt{n})$ for it to be the (size unrestricted) densest subgraph of M.

Approximation algorithms

There is a vast literature on approximation algorithms for NP-hard problems. In the cases of interest to us however the best approximation guarantees available are too weak and are of a different flavor from results we would ideally like to have. Consider for instance the densest k- subgraph problem. This problem is NP-hard and does not admit a constant factor approximation algorithm unless NP has sub-exponential time algorithms. The best known approximation guarantee was recently obtained in [4]. They show that given any graph on n vertices there is an algorithm that runs in time $O(n^{O(1/\epsilon)})$ which will find a k-subgraph of density $\Omega(d_{\max}/n^{1/4+\epsilon})$, where $d_{\max}$ is the density of the densest k-subgraph, i.e. for any constant ϵ there is a polynomial time algorithm that achieves an approximation ratio of $O(n^{1/4+\epsilon})$.

The result however is not good enough for us to make a meaningful statistical guarantee. We would like a guarantee that says that we recover the true bicluster (or something that is close to it in Hamming distance) with high probability. At the minimax scaling, we know that the densest 2k-subgraph is the bicluster we are interested in. However, it can be shown that at this scaling there are several other $(k \times k)$ submatrices whose density is within a constant factor of the density of the true bicluster. This means that for any approximation ratio worse than $O(1)$, the approximation algorithm could find a $(k \times k)$ submatrix that might not overlap *at all* with the true submatrix. For the biclustering problem we are not interested in good approximation guarantees on the objective function, which we use only as a proxy. Rather, we are interested in guaranteeing that with high probability the submatrix we find is close to the true submatrix of activation.

5 Conclusions

The biclustering problem highlights the tradeoff between computational complexity and statistical efficiency. The most significant open question with respect to our work is: "Is there a computationally efficient algorithm that achieves the minimax rate for all tuples (n, k, μ)?."

While we conjecture that the biclustering problem is computationally hard, the structure and randomness pose significant obstacles to the direct application of reductions to show hardness. In the biclustering problem we are given a particular *structured, random* (and *not* arbitrary, worst-case) instance of a known NP-hard problem. Showing that even these seemingly benign instances are not significantly easier than the worst-case instances is an important direction for future work.

Our work also highlights an important shortcoming of minimax analysis with regards to computational tractability. Ideally rather than being defined as an infimum over *all* estimators we would like to be able to define the minimax rate over a smaller class of all *efficiently* computable estimators, and develop tools to study this restricted minimax rate. Formalizing this notion is also an important direction of future work.

References

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

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

[3] 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.

[4] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities – an O($n^{1/4}$) approximation for densest k-subgraph. *CoRR*, abs/1001.2891, 2010.

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

[6] Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In *APPROX*, pages 84–95, 2000.

[7] A. d Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434, 2007.

[8] Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, and Aarti Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems 2011*. Accepted for publication, 2011.

[9] 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.

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

[11] 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.

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