Yale University
Department of Computer Science

Spectral Clustering with Limited Independence

Anirban Dasgupta
Cornell University

John Hopcroft
Cornell University

Ravi Kannan
Yale University

Pradipta Mitra
Yale University

YALEU/DCS/TR-1340
December 6, 2005
Updated September 5, 2006

Pradipta Mitra is supported by NSF's ITR program under grant number 0331548. Ravi Kannan is supported by NSF Award CCF 0310805

Abstract

This paper considers the well-studied problem of clustering a set of objects under a probabilistic model of data in which each object is represented as a vector over the set of features, and there are only k different types of objects. In general, earlier results (mixture models and "planted" problems on graphs) often assumed that all coordinates of all objects are independent random variables. They then appeal to the theory of random matrices in order to infer spectral properties of the *feature* × *object* matrix. However, in most practical applications, assuming full independence is not realistic.

Instead, we only assume that the objects are independent, but the coordinates of each object may not be. We first generalize the required results for random matrices to this case of limited independence using some new techniques developed in Functional Analysis. Surprisingly, we are able to prove results that are quite similar to the fully independent case modulo an extra logarithmic factor. Using these bounds, we develop clustering algorithms for the more general mixture models. Our clustering algorithms have a substantially different and perhaps simpler "clean-up" phase than known algorithms. We show that our model subsumes not only the planted partition random graph models, but also another set of models under which there is a body of clustering algorithms, namely the Gaussian and log-concave mixture models.

1 Introduction

In a wide range of applications, one analyzes a collection of m objects, each of which is a vector in n-space. The input consists of a $n \times m$ matrix A, each column representing an object and each row representing a "feature". An entry of the matrix then stands for the numerical value of an object feature. Term-Document matrices (where the entries may stand for the number of occurrences of a term in a document) and product-customer matrices (where the entries stand for the amount of a product purchased by a customer) are two salient examples. An important question regarding such data matrices widely analyzed in Data Mining, Information Retrieval and other fields is this:

Assuming a probabilistic model from which each object is chosen independently, can one infer the model from the data?

In an easier version of the above question, we set out by assuming that the probabilistic model is a mixture[1] of k simple distributions (where k is very small, in particular, $k \ll m, n$). What is then the required condition on the probability distributions so that we can group the objects into k clusters, where each cluster consists precisely of objects picked according to one of the component distributions of the mixture?

There has been much success in this and also in the so-called "Planted Partition" models, but only under a restrictive assumption that all the entries of the matrix A are **independent**. We will refer to this as the **full independence assumption**. Indeed, the work of Azar, Fiat, Karlin, McSherry and Saia [5] formulates the above questions (starting with similar examples) and tackles the problem under the full independence assumption. The method has received considerable attention in the planted partition graph models as well [7, 2, 3, 11]. More directly relevant to this paper are the papers [19, 8], which show that assuming full independence and certain separation conditions between the means of the component distributions, the projection of objects to the space spanned by the top k singular vectors of A leads to a clustering based just on distances which clusters most objects correctly. Then, as in earlier algorithms, there is a "clean-up" phase which now correctly clusters all the objects. The clean-up phase turns out to be technically complicated.

The general reason for the full independence assumption is that with this in hand, one may rely on the theory of random matrices, initiated by Wigner [22, 23]. The central result of this theory is that for an $n \times n$ symmetric matrix X, whose entries in the upper triangular part are **independent random variables** with mean 0, there are very good upper bounds on the largest eigenvalue (as proved by Füredi and Komlos [15] and Vu [25]). Indeed, such bounds play a crucial role in all the spectral algorithms in answering the above questions. Let A denote the input matrix generated by a random process that has k component distributions and $\bar{A}$ denote the expectation of A. The matrix concentration result applied to the matrix $X = A - \bar{A}$, the difference between A and its expectation, implies that the span of the top k eigenvectors of A is very close to the span of the k centers of the k component distributions. This in turn says that after projecting the data onto the span of these k-eigenvectors, a simple clustering algorithm recovers almost all points in the original clusters.

Another direction in learning mixture models, and, one that has been more successful in dealing with dependence, is to assume the distributions to be general Gaussians or log-concave in which the coordinates of one object can depend on one another. While there has been quite some success in this model [9, 4, 24, 17, 1], the spectral algorithm in these papers work mainly for Gaussians and/or

[1] A mixture of k probability densities $P_1, P_2, \ldots P_k$ is a density of the form $w_1P_1 + w_2P_2 + \ldots + w_kP_k$, where the w_i are fixed non-negative reals summing to 1.

log-concave distributions. (But the inter-center separation one needs to assume for Gaussians is better than what we will get here in the general case.) This also means that this line of work does not include the case of the planted partition graph models, where the distributions tend to be discrete. A main observation we use in this paper is that in many discrete distributions, one gets tight concentration around the mean only when a lot of coordinates are considered, ie, when the direction being considered is "balanced" (we make these notions precise later).

The point of departure of this paper is that the assumption of full independence is probably the most important impediment to applying these results in practice; indeed in the two examples mentioned above, while one may assume that the documents are independent of each other, it is certainly not true that the occurrence of different terms inside a document are independent (the same thing can be said about the product-customer case). Indeed the generative model of Papadimitrou, Raghavan, Tamaki and Vempala [20] illustrates this point well. Similarly, in the product-customer model, while different customers may be reasonably assumed to be independent of each other, one customer does not choose various products independently. At the minimum, the customer is subject to budget constraints. The main contribution of this paper will be to replace the full independence assumption with a **limited independence assumption**, namely, positing that the objects are independent, but the features may not be. (So the columns of A are independent, but the rows need not be.)

Our model in this paper starts out by assuming that the input points are samples from a probability distribution that satisfies certain concentration properties, and obeys the limited independence assumption. The concentration properties that we assume are general enough to encompass both the planted partition and the log-concave distribution models. Under these set of assumptions, we solve the mixture learning problem using spectral methods. The most important ingredient in our proofs is the matrix inequality of Theorem 6 and Corollary 7 that may be of independent interest in the theory of random matrices - namely, we prove a bound on the spectral norm of (possibly rectangular) matrices under limited independence. Surprisingly, the bounds are similar to the ones proved under full independence, except for logarithmic (in n) factors. The separation conditions that we require for our algorithm are similar to the best known results in the planted partition model by [19] and differs by logarithmic factors from the best results in the log-concave distribution model [17, 1].

The general method used for proving bounds under full independence originated with Wigner's work - it consisted of bounding the trace of a high power of the square matrix. Our matrices being rectangular, such an approach cannot be carried out in a simple fashion. Instead, we rely on certain techniques recently developed in Functional Analysis (see Rudelson in [21]) to prove our theorem. These techniques were developed as a means to solving a different problem - namely, what is the minimum number of independent identically distributed samples from a n-dimensional Gaussian density with the property that the variance-covariance matrix of the samples approximates the variance-covariance matrix of the actual density to within small relative error? The result presented here is similar to that presented in [21], but as the higher moment bound we require is not easily recoverable from the results there (which are proved in the context of vectors in isotropic positions), we provide a separate proof and cast our result in a way easily applicable to clustering problems.

On the algorithmic side, the basic difference of our technique with earlier algorithms is in the cleanup phase. The cleanup phase of earlier algorithms was either easy, as in the case of log-concave distributions [17, 24], or had to be done by constructing a combinatorial projection for the planted partition case, as in [19, 8]. The construction of the combinatorial projection in the planted partition case exploits the fact that since we are dealing with graphs, both rows and columns of the matrix represent vertices and so one can alternately cluster rows or columns. This symmetry is not present in our general model; here we are able to cluster just the objects using the features, but not cluster the features themselves. Because of this and the dependency of the coordinates, cleaning up the solution appears to be technically a much harder task. The cleanup phase constructed in this paper is, on the contrary, arguably simpler than the previous constructions of combinatorial projections [19, 8].

2 Model and Result

We start with a set of k distributions $P_1, P_2 \dots P_k$ in $\mathbb{R}^n$. The center (i.e. expectation) of the r^{th} distribution is denoted by μ_r. There is also a set of mixing weights $\{w_r \mid r = 1 \dots k\}$ associated with the k distributions, such that $\sum_r w_r = 1$ and each $w_r = \Omega(\frac{1}{k})$. The minimum of the weights is denoted as $w_{\min}$. The data is generated as follows. In generating the i^{th} sample, denoted as A_i, we first pick a distribution, say P_r with probability w_r. Then, the sample A_i is chosen from distribution P_r independently from all the other samples. The m samples A_i form the columns of the input matrix $A \in \mathbb{R}^{n \times m}$. Thus, the columns of A are chosen independently, while there is no assumption on the independence on the coordinates of each sample A_i. The expectation of the sample A_i is denoted as $\mathbf{E}[A_i] = \bar{A}_i$. An important notion for us is the definition of *balance* of a vector. Define the balance of a vector v as $\eta(v) = \frac{\|v\|_2}{\|v\|_\infty}$. Intuitively, $\eta(v)$ indicates the number of "significant" coordinates in v. Note that for all vectors $v \in \mathbb{R}^n$, $1 \le \eta(v) \le \sqrt{n}$. In this paper, we will use $poly(m, n)$ to be a polynomial in m and n with suitably large coefficients.

The concentration result we derive in section 5 is quite general, hence the form of our assumptions and conditions presented below are just an instantiation of the wide range of results that we actually achieve. Our presentation of the bounds in terms of $\sigma\sqrt{\log m}$ is motivated by the interesting case of $m = n$ where Chernoff type results give $\sigma\sqrt{\log n}$ bounds for many natural problems. We will assume $m \ge n$ without loss of generality to avoid cumbersome expressions like $\max(\sqrt{\log m}, \sqrt{\log n})$ in our bounds.

Mixture Model. The following are our required conditions for each of the probability distributions.

1. Maximum variance of any P_r in any direction is at most σ^2. That is, for any vector v of unit length, we have that $\mathbf{E}\left[(v \cdot (A_i - \bar{A}_i))^2\right] \le \sigma^2$.

2. There exists $\eta^* \in [1, \sqrt{n}]$ such that for each **fixed** unit vector v that has balance at least η^*, and for each sample A_i,

$$\mathbf{Pr}(|v \cdot (A_i - \bar{A}_i)| \ge \sigma\sqrt{\log m}) \le \frac{1}{poly(n, m)}. \tag{1}$$

An orthonormal basis of balance $\Theta(\sqrt{n})$ can be found for a n-dimensional space[2]. This, along with the previous condition implies that, for each sample A_i, we have the following bound on

[2] This can be seen by finding a basis for the space $\mathbb{R}^n$ that consists of vectors from $\{-1, 1\}^n$ only. Standard results [6, 16] show that such a basis, known as Hadamard basis, must exist if the dimensionality n is a multiple of 4. We can increase the dimensionality n to be a multiple of 4, without losing anything on the separation conditions, and losing only a constant factor in balance.

the deviation $\left\|A_i - \bar{A}_i\right\|$, i.e.

$$\mathbf{Pr}(\|A_i - \bar{A}_i\|_2 \geq \sigma\sqrt{n\log m}) \leq \frac{1}{poly(n,m)}. \tag{2}$$

Remark 1. *For Gaussian distributions η^* can be as small as $O(1)$ whereas for independent $0/1$ distributions, we need $\eta^* = \Theta(\sqrt{n})$ for condition 2 to hold.*

Separation Condition. We assume the following about the centers of the distribution.

1. The distributions are separated in the sense that for each pair of distributions P_r and P_s with corresponding centers μ_r and μ_s, and for a large enough constant c we have that,

$$\|\mu_r - \mu_s\|_2 \geq 40c\sigma k\sqrt{\frac{\log(k)}{w_{\min}}}\left(\sqrt{\log m} + \sqrt{\frac{n}{m}\log n\log m} + 1\right).$$

2. All the pairwise difference vectors of centers i.e. all vectors $\mu_r - \mu_s$ for all r and s should have balance at least $\min(2\eta^*, \sqrt{n})$ where η^* is the balance requirement of the probability distributions. (Having balanced centers is perhaps a more natural assumption, but note that ours is a generalization of that).

For brevity, we will define τ as follows,

$$\tau = c\sigma k\sqrt{\frac{\log(k)}{w_{\min}}}\left(\sqrt{\log m} + \sqrt{\frac{n}{m}\log m\log n} + 1\right). \tag{3}$$

2.1 Result

The following is the main result in our paper.

Theorem 1. *Given m samples taken from the mixture of k distributions that satisfy the above conditions, there is an algorithm that given the values of σ, k and η^*, classifies all the samples correctly with probability $1 - \frac{k}{m} - \frac{1}{n^2}$ over the input distribution, and probability $1 - \frac{1}{4k}$ over the random bits of the algorithm.*

Relation to previous work An important subcase of our framework is the planted partition model. Our framework does not apply to the models for symmetric random graphs due to the symmetry requirements. However, the models for directed graphs can be viewed in this "samples from a mixture distribution" framework, as we can view each vertex as a sample from a component distribution P_i, and choose its vector of outgoing edges according to P_i. Our required separation between the centers is then similar to the best known result of [19]. The balance requirement η^* of the distributions is akin to saying that each cluster be large enough. When the probability distributions P_r are Gaussians or log-concave distributions, the balance condition is basically unnecessary, and the pairwise separation between centers that we require is worse off than Kannan et al. [17] by a factor of $\log n$.

Our conditions for convergence of distributions is similar to the condition of *f-convergence* posed by Achlioptas and McSherry in [1]. However, the important distinction is that [1] requires convergence in *every* direction. This property is not satisfied by many discrete distributions over $\{0,1\}^n$.

By relaxing the requirement to being valid for balanced directions only, we can encompass the planted partition graph models and other such discrete models. The following example illustrates the necessity of the balance condition for such distributions for our algorithm.

Example. Suppose both P_1 and P_2 are probability distributions over $\{0,1\}^n$ with centers $\mu_1 = (p,\ldots,p)$ and $\mu_2 = (Cp,p,\ldots,p)$, for any C such that $Cp < 1$. Each sample A_i from P_r is generated by choosing each coordinate A_{ij} independently to be one with probability μ_{rj} and zero otherwise. The distributions are separated along the direction $(1,0,\ldots,0)$ which is not a balanced direction. Furthermore, since the distributions are different only along the first coordinate, even with knowledge of the centers, we are bound to misclassify $\Theta(1-Cp)$ fraction of the points.

3 Algorithm

In this section, we present our algorithm to separate mixtures of distributions. The algorithm needs an estimate of the separation between clusters, the balance η^*, and the knowledge of k, the number of clusters into which to partition the set of samples.

The main idea of the algorithm **Cluster** is as follows. We really would like to find the subspace spanned by the k centers $\{\mu_r\}$ and project all points onto that space. This projection would preserve the distance between the centers and concentrate all the samples around their corresponding expectations. Unfortunately, we do not know this subspace. Instead, we take the spectral rank-k approximation to the data, and get a subspace that is close to the expected subspace. Then we cluster the rank-k representations $A_i^{(k)}$, that are the projections of points A_i onto this subspace. A large (but constant) fraction of the points are now correctly classified. At this point, in order do cleanup and classify the rest of the points, different techniques have been employed for different mixture models, none of which can be applied to our model.

The misclassification error occurs because the subspace spanned by the singular vectors might not be balanced and hence the points might not be concentrated around their respective centers. The same is true for the subspace spanned by the approximate centers obtained from the first stage. To overcome this difficulty, we draw a set of $\binom{k}{2}$ lines through the pairs of approximate centers. We then use a "smoothening" procedure, the subroutine **Balance**, to find a set of balanced lines that are close to each of these lines. Projecting onto each of the balanced lines results in the points being concentrated around their centers, and the distances between corresponding centers being preserved. Points belonging to a certain cluster will be close to the corresponding center on each of $k-1$ lines that pass through that center. In order to decondition the construction of the projection vectors from the actual classification, the algorithm uses two sets of samples A and B. The **Balance** subroutine takes in a vector v' and an error measure ε and tries to find a vector $\tilde{v}$ that is ε-close to v' and has a good balance.

4 Proofs for General Separation

Before stating the actual lemmas, we motivate the broad picture. The rank-k approximation of a matrix A is denoted as $A^{(k)}$, and projection matrices are denoted by π (suitably subscripted). First, we will show that for distributions that obey the stated assumptions, $A^{(k)}$ is indeed a close approximation to the expectation matrix $\bar{A}$. In doing so, we have to prove that the error matrix

Algorithm 1 Balance(v', ε)

```
1: Sort the entries of $v'$ in absolute value, say $|v'_{i_1}| \geq \ldots \geq |v'_{i_n}|$, and pick the $t$ such that $\sum_{j \leq t-1}(|v'_{i_j}| - |v'_{i_t}|)^2 < \varepsilon^2$,   $\sum_{j \leq t}(|v'_{i_j} - v'_{i_{t+1}}|)^2 \geq \varepsilon^2$. (easily done by binary search)
2: Now find $|v'_{i_t}| \geq a \geq |v'_{i_{t+1}}|$ such that $\sum_{j \leq t-1}(|v'_{i_j}| - a)^2 = \varepsilon$ (using any standard numerical algorithm)
3: Return $sign(v'_{i_1})a, \ldots sign(v'_{i_t})a, v'_{t_{t+1}}, \ldots v'_{t_n}$ where sign is $1/-1$ depending on whether $v'_{i_1}$ is positive or negative.
```

Algorithm 2 Cluster(S, τ, k)

```
1: Randomly divide the set of samples into two sets $A$ and $B$.
2: Find $A^{(k)}$, the rank-$k$ approximation of the matrix $A$.
3: Find out a set of pseudo-centers from the (initially all unmarked) columns of $A^{(k)}$ using the following method.
     a. Randomly choose an unmarked column $i$ as a new pseudo-center.
     b. For all columns $j$ such that $\|A_i^{(k)} - A_j^{(k)}\| \leq 2\tau$, mark the column $j$.
     c. Continue the previous steps (a)-(b) till we get $k$ centers or till there are at most $w_{\min}m/4$ columns left to mark that are not assigned anywhere.
4: Call each of the $l$ columns $A_{i_1}, \ldots, A_{i_l}$ chosen in step (b) to be the $l$ center estimates $\mu'_1, \ldots, \mu'_l$.
5: For each of the pair of centers $r, s$ construct $v'_{rs} = \mu'_r - \mu'_s$.
6: We will correct the balance of each difference vector $v'_{rs}$. Individually balance all the vectors $v'_{rs}$ by invoking $\tilde{v}_{rs} = \mathbf{Balance}(v'_{rs}, 2\tau)$.
7: For all $r, s$, project each center $\mu'_r$ and $\mu'_s$ onto $\tilde{v}_{rs}$. Then project each $B_i \in B$ onto each of the vectors $\{\tilde{v}_{rs}\}$ and classify it as belonging to either cluster $r$ or cluster $s$ depending on whether it is close to the projection of $\mu'_r$ or $\mu'_s$ on $\tilde{v}_{rs}$.
8: A sample $B_i$ is classified finally as belonging to cluster $r$ if it is classified under $r$ in the all tests $\{v'_{rs}\}$ (for every $s$).
```

$A - \bar{A}$ has a small 2-norm, which indicates that the errors $A_i - \bar{A}_i$ cannot mislead the search for the "best rank-k" subspace. This is done in Lemma 2.

Once we establish this closeness, it will follow that the center estimates that we construct are close approximations to the original centers. That is, we will show that in the steps (2a)-(2c) of the algorithm **Cluster**, we create k center estimates $\mu'_1, \ldots, \mu'_k$, one for each distribution, and each of them is not very far off from the corresponding μ_r. Unfortunately, this is not enough to show that we can label all the points correctly. We still have to guarantee that *balance* of the subspace spanned by the approximate centers $\{\mu'_1, \ldots, \mu'_k\}$ is at least η^* so that the samples are concentrated around their expectations upon projection to this space. This is shown in Lemma 4 and corollary 5. These lemmas lead the the final proof of Theorem 1.

Lemma 2. *Under the stated assumptions about the probability distribution, with probability* $1 -$

$\frac{1}{poly(m,n)} - \frac{1}{n^2}$, *the rank-$k$ approximation matrix $A^{(k)}$ satisfies*

$$\|A^{(k)} - \bar{A}\|_F \leq c\sigma\sqrt{k}\left(\sqrt{m} + \sqrt{n \log m \log n}\right).$$

where c is a large enough constant.

Proof. This lemma follows by a standard argument from corollary 7. Corollary 7 is a special case of a concentration result of independent interest, the proof of which will be presented in Section 5. The proof of Lemma 2 itself is in the appendix. □

The proof of the following lemma is akin to similar results from [19] and [8]. Note that at this point we are interested not in a complete clustering, but in choosing good centers only.

Lemma 3. *Given the matrix bounds from Lemma 2, with probability $1 - \frac{1}{4k}$, we choose only k columns in step (2a) of **Cluster**. Further, the k columns $\{A^{(k)}_{i_r} \mid r = 1 \ldots k\}$ chosen are each from different clusters, and each satisfies $\|A^{(k)}_{i_r} - \mu_{i_r}\| \leq \tau$ where μ_{i_r} is the center of the cluster that column i_r belongs to.*

Proof. See appendix. □

We now show that the **Balance** algorithm actually balances each vector $v'_{rs} = \mu'_r - \mu'_s$.

Lemma 4. *Suppose we are given a vector v' with $\|v'\| > 20\tau$. Then, if $\tilde{v} =$**Balance**$(v', 2\tau)$, and x be such that $\|x - v'\|_2 \leq 2\tau$, then $\eta(x) \leq 2\eta(\tilde{v})\frac{\|v'\|+2\tau}{\|v'\|-2\tau} \leq 2\eta(\tilde{v})$.*

Proof. Let us assume wlog that all vectors involved have only positive entries. Also, assume wlog that the indices in both v' and x are sorted according to the same order, i.e. $v'_1 \geq v'_2 \ldots$ and $x_1 \geq x_2 \ldots$.

First we claim that for any such x, $\|x\|_\infty \geq \|\tilde{v}\|_\infty$. It is clear that $x_1 = \|x\|_\infty$. If $x_1 < \tilde{v}_1$, then $x_i < \tilde{v}_i$ for $i \leq t$ (t is the index found in the algorithm **Balance**), and it is clear that $\|v' - x\| > \|v' - \tilde{v}\| = 2\tau$. This is a contradiction.

Now, $\|\tilde{v}\|_2 \geq \|v'\|_2 - 2\tau$ and $\|x\|_2 \leq \|v'\|_2 + 2\tau$. Hence, $\eta(x) = \frac{\|x\|_2}{\|x\|_\infty} \leq \frac{\|x\|_2}{\|\tilde{v}\|_\infty} \leq \frac{\|\tilde{v}\|_2}{\|\tilde{v}\|_\infty}\frac{\|x\|_2}{\|\tilde{v}\|_2} \leq \eta(\tilde{v})\frac{\|v'\|+2\tau}{\|v'\|-2\tau} \leq 2\eta(\tilde{v})$ □

Corollary 5. *For each pair of center r, s found in the step (3) of the algorithm, the vector $\tilde{v}_{rs} =$ **Balance**$(v'_{rs}, 2\tau)$ satisfies $\eta(\tilde{v}_{rs}) \geq \eta^*$.*

Proof. Let $v_{rs} = \mu_r - \mu_s$. Now from lemma 3, $\|\mu_r - \mu'_r\| \leq \tau$. Clearly $\|v_{rs} - v'_{rs}\| \leq 2\tau$. Now by the balance condition stated in Section 2, v_{rs} has balance at least $2\eta^*$, and invoking lemma 4, we get that $\tilde{v}_{rs}$ has balance at least η^*. □

Thus finally, we can prove the theorem 1.

Proof. We first give a sketch of the proof. By the results of Lemma 3, we know that each of the k approximate centers μ'_r is not too far from the actual center μ_r. We also know that after balancing, each vector $\tilde{v}_{rs}$ is at most 2τ distant from the vector v'_{rs}. Thus, we can show that the difference between the approximate centers μ'_r and μ'_s will be well preserved on projection to $\tilde{v}_{rs}$. Also, by virtue of balancing, each sample will be close to its expectation upon projection to $\tilde{v}_{rs}$.

Thus, projecting all samples onto $\tilde{v}_{rs}$ and using the projection of the centers μ'_r and μ'_s to label the points, the points that are actually from P_r and P_s are labeled correctly. For a single sample B_i from P_r, testing for all $\binom{k}{2}$ projections, pairwise comparisons will reveal the actual pseudo-center r.

Here are the details. For each r, s, let the projection matrix onto $\tilde{v}_{rs}$ be denoted by $\tilde{\pi}_{rs} = \frac{\tilde{v}_{rs}\tilde{v}_{rs}^T}{\|\tilde{v}_{rs}\|^2}$. The projection on v'_{rs} is similarly denoted as π'_{rs}. The algorithm projects each sample B_i onto the vector $\tilde{v}_{rs}$ and classifies it as belonging to distribution r or s depending on whether it is closer to $\tilde{\pi}_{rs}\mu'_r$ or $\tilde{\pi}_{rs}\mu'_s$. We first show that the projection of the approximate centers are separated.

$$\begin{aligned}\|\tilde{\pi}_{rs}(\mu'_r - \mu'_s)\| &= \|(\pi'_{rs} + (\tilde{\pi}_{rs} - \pi'_{rs}))(\mu'_r - \mu'_s)\| \\ &\geq \|\pi'_{rs}(\mu'_r - \mu'_s)\| - \|(\tilde{\pi}_{rs} - \pi'_{rs})(\mu'_r - \mu'_s)\| \\ &\geq \|\mu'_r - \mu'_s\| - \|\tilde{\pi}_{rs} - \pi'_{rs}\|\|\mu'_r - \mu'_s\| \end{aligned} \tag{4}$$

Using a simple consequence of of Stewart's theorem (see Fact 13 in the Appendix), $\|\pi'_{rs} - \tilde{\pi}_{rs}\| \leq \frac{\|\tilde{v}_{rs} - v'_{rs}\|}{\|\tilde{v}_{rs}\| - \|\tilde{v}_{rs} - v'_{rs}\|} \leq \frac{2\tau}{10\tau - 2\tau} \leq \frac{1}{4}$. Employing this in equation (4), and noting that $\|\mu'_r - \mu_r\|$ is small,

$$\begin{aligned}\|\tilde{\pi}_{rs}(\mu'_r - \mu'_s)\| \geq \|\mu'_r - \mu'_s\| - \|\tilde{\pi}_{rs} - \pi'_{rs}\|\|\mu'_r - \mu'_s\| &\geq \frac{3}{4}\|\mu'_r - \mu'_s\| \\ &\geq \frac{3}{4}\left(\|\mu_r - \mu_s\| - \|\mu_r - \mu'_r\| - \|\mu_s - \mu'_s\|\right) \\ &\geq \frac{3}{4}(40\tau - \tau - \tau) \geq 20\tau \end{aligned}$$

Because each of the vectors $\tilde{v}_{rs}$ is η^*-balanced, we have that, for each sample B_i in B, with probability $1 - \frac{1}{poly(n,m)}$, $\|\pi_{rs}(B_i - \mathbf{E}[B_i])\| \leq \sigma\sqrt{\log m}$. Thus, if the sample B_i is from the distribution P_r, then the distance of $\pi_{rs}(B_i)$ from the projected center estimate $\pi_{rs}(\mu'_r)$ is at most

$$\begin{aligned}\|\pi_{rs}(B_i - \mu'_r)\| &\leq \|\pi_{rs}(B_i) - \pi_{rs}(\mathbf{E}[B_i])\| + \|\pi_{rs}(\mathbf{E}[B_i]) - \pi_{rs}(\mu'_r)\| \\ &\leq \sigma\sqrt{\log m} + \|\pi_{rs}(\mu_r - \mu'_r)\| \leq \sigma\sqrt{\log m} + 2\tau \end{aligned}$$

Also, the distance of $\pi_{rs}(B_i)$ from the other projected center estimate $\pi_{rs}(\mu'_s)$ is at least

$$\begin{aligned}\|\pi_{rs}(B_i - \mu'_s)\| &\geq \|\pi_{rs}(\mathbf{E}[B_i]) - \pi_{rs}(\mu'_r)\| - \|\pi_{rs}(B_i) - \pi_{rs}(\mathbf{E}[B_i])\| \\ &\geq \|\pi_{rs}(\mu_s - \mu'_r)\| - \sigma\sqrt{\log m} \\ &\geq \|\pi_{rs}(\mu'_s - \mu'_r)\| - \|\pi_{rs}(\mu_s - \mu'_s)\| - \sigma\sqrt{\log m} \\ &\geq 20\tau - 2\tau - \sigma\sqrt{\log n \log m} \end{aligned}$$

Thus, $\|\pi_{rs}(B_i - \mu'_r)\| < \|\pi_{rs}(B_i - \mu'_s)\|$, and hence, in the test that involves projection onto $\tilde{v}_{rs}$, each sample B_i that belongs to P_r is actually classified under P_r and each sample B_j belonging to P_s is actually classified under P_s. Any sample belonging to other clusters may be classified under either one of them. Thus, for each sample B_j, only one center μ'_{r_j} beats all the other centers in pairwise tests, and hence this is the actual cluster that B_j belongs to. The probability of correctness of the algorithm is controlled by the following factors. The random matrix bound in Corollary 7 holds with probability $1 - \frac{m}{poly(n,m)} - \frac{1}{n^2}$. As per Lemma 3 the greedy clustering on the columns of $A^{(k)}$ gives us a good set of centers with probability $1 - \frac{1}{4k}$. All the projections of the m samples on $\binom{k}{2}$

balanced vectors are all concentrated with probability $1 - \frac{mk^2}{poly(n,m)}$. Thus the total probability of success over the random matrix model is at least $1 - \frac{k^2}{n} - \frac{1}{n^2}$ and the (boostable) probability of success of the random bits of the algorithm is $1 - \frac{1}{4k}$. □

Remark. Our clean-up phase is quite different compared to previous work in planted partition model[8, 19]. The main paradigmatic changes are the following. We move from projecting on a k-dimensional subspace to a number of one dimensional subspaces, and thereby avoiding so-called "combinatorial projections", which implicitly needed the fact that feature-space is clusterable (true in the graph models, as "objects" and "features" are the same, they are vertices). We also avoid upper bounding $\|\pi(\mu_r) - \mu_r\|$ (Here, π is whatever the relevant projection is), and rather lower bound $\|\pi(\mu_r - \mu_s)\|, r \neq s$. The lower bound is implied by the earlier upper bound, and hence often easier to prove and applicable in a wider range of situations.

5 A Concentration Result

In this section we prove the concentration result on the spectra of the random matrix A. We prove a general bound on the matrix AA^T from which the result on the norm of $A - \bar{A}$ follows. Note that $AA^T = \sum_{i=1}^m A_i A_i^T$, where each $A_i A_i^T \in \mathbb{R}^{n \times n}$. We denote $\mathbf{E}\left[AA^T\right] = D$. We will actually bound a high moment of $||AA^T||$, i.e., we will bound $\mathbf{E}_A||AA^T||^l$ for any even positive integer l.

Theorem 6. *For any even $l > 0$, we have $\mathbf{E}_A||AA^T||^l \leq 2^{l+2}||D||^l + 2^{4l+2}n^2 l^{l+4}\mathbf{E}_A[\max_i |A_i|^{2l}]$.*

Before proving this theorem, we first give a corollary of the theorem that is useful to us. We apply the theorem to $A - \bar{A}$ (instead of to A). Note that $\mathbf{E}\left[(A - \bar{A})(A - \bar{A})^T\right] = \mathbf{E}\left[AA^T\right] - \bar{A}\bar{A}^T$. Recall that the maximum variance of any A_i in any direction is at most σ^2. Then it is easy to see that for any unit length vector v, $v^T(\mathbf{E}\left[AA^T\right] - \bar{A}\bar{A}^T)v = \sum_i \mathbf{E}\left[(v^T A_i)^2\right] - (\mathbf{E}\left[v^T A\right])^2 \leq m\sigma^2$. So, $\|\mathbf{E}\left[AA^T\right] - \bar{A}\bar{A}^T\| \leq m\sigma^2$; this bounds the $\|D\|$ term of the theorem. For bounding $\max_i \|A_i - \mathbf{E}\left[A_i\right]\|$, recall that for all i, $\|A_i - \bar{A}_i\| \leq \sigma\sqrt{n \log m}$ with probability $1 - \frac{1}{poly(n,m)}$.

Now, we apply the Theorem (with $l = \log n$) to $A - \bar{A}$ to get :

Corollary 7. *Under the above conditions, we have for all $\theta > 0$,*

$$\mathbf{Pr}\left(\|A - \bar{A}\| \geq \theta\sigma(4\sqrt{m} + 100\sqrt{n \log n \log m})\right) \leq \frac{1}{poly(m,n)} + \frac{1}{n^{\log\theta - 8}}.$$

Proof. (of the Theorem) We pick an auxiliary set of random vectors $B_1, B_2, \ldots B_m$, where for each i, B_i has the same distribution as A_i; the B_i form the matrix B, say. We also pick another set of auxiliary random variables - $\zeta_1, \zeta_2, \ldots \zeta_m$ where $A, B, \zeta_1, \zeta_2, \ldots \zeta_m$ are all independent and each ζ_i is ± 1 with probability $1/2$. We let $\zeta = (\zeta_1, \zeta_2, \ldots \zeta_m)$. Let $p(A)$ denote the probability (or probability density) of a particular A. We allow discrete as well as continuous distributions, but we will use integral for both (and not bother to use sums for discrete distributions). Note that $p(\cdot)$ induces a probability measure on AA^T - say $q(AA^T)$ and that $\mathbf{E}_p(AA^T) = \mathbf{E}_q(AA^T)$. The starting point of this proof (like many other proofs on eigenvalues of random matrices) is to observe that for a symmetric matrix X, $\|X\|^l \leq \mathrm{Tr}(X^l)$. It is the trace of a power that we bound for most of the proof. We will use two well-known facts stated below. (See for example, [6], IV.31).

Proposition 1. *For any even integer l, $(Tr(X^l))^{1/l}$ is a norm (called a Schatten norm). Hence it is a convex function of the entries of the matrix X and thus so is $Tr(X^l)$. Also, we have for any two matrices X, Y, $(Tr((X+Y)^l))^{1/l} \le (Tr(X^l))^{1/l} + (Tr(Y^l))^{1/l}$.*

It is also easy to see that $\mathrm{Tr}(X^l) \le n\|X\|^l$ (which means that for $l \ge \log n$, the 2-norm and the Schatten norm are in fact *equivalent*). We will need the following two lemmas:

Lemma 8. $\mathbf{E}_A\, Tr((AA^T - D)^l) \le 2^{l+1}\mathbf{E}_A\mathbf{E}_\zeta\, Tr\left(\left(\sum_{i=1}^m \zeta_i A_i A_i^T\right)^l\right)$.

Proof. See appendix. □

Lemma 9. *There is a constant c such that, for each fixed A, we have*

$$\mathbf{E}_\zeta\, Tr\left(\left(\sum_i \zeta_i A_i A_i^T\right)^l\right) \le l^{l/2}\max_i |A_i|^l\, Tr\left(\left(\sum_i A_i A_i^T\right)^l\right)^{\frac{1}{2}}.$$

Proof. This result is essentially proved in [21] (in the required higher moment form, see equ. (3.4), pp 66). We will omit details here. □

Using the two lemmas, and noting the relation between the two norms,

$$\mathbf{E}_A\|AA^T\|^l \le 2^l\|D\|^l + 2^l E_A\|AA^T - D\|^l \le 2^l\|D\|^l + 2^{2l+1}nl^{(l/2)+1}\mathbf{E}_A\left(\max_i \|A_i\|^l\|AA^T\|^{l/2}\right)$$
$$\le 2^l\|D\|^l + 2^{2l+1}nl^{(l/2)+1}\left(\mathbf{E}_A\max_i\|A_i\|^{2l}\right)^{1/2}\left(\mathbf{E}_A\|AA^T\|^l\right)^{1/2}.$$

Letting $X = \sqrt{\mathbf{E}_A\|AA^T\|^l}$, the above gives a quadratic inequality for X; it is easy to see that the inequality implies that X is at most the larger of its roots. This implies the Theorem.

6 Conclusion

A number of natural open questions arise from our work. One motivation for norm concentration results in random matrix theory has been that these results are related to expansion properties in graphs. Our theorem on matrices with limited dependence, however, does not imply a very strong expansion property due to the extra logarithmic factors involved. It is an important question whether these bounds can be strengthened further. Another interesting direction is to see whether we can extend this framework to the learning of heavy tailed mixture models.

Besides the clustering problem, we may also consider a related problem in collaborative filtering and matrix reconstruction [5, 10]: here one has some entries of say the product-customer matrix A and have to infer the whole matrix assuming A was low-rank and possibly also a generative(probabilistic) model for A. There again results are for full independence. We believe our work can be extended to tackle these problems under a limited independence assumption.

References

[1] Dimitris Achlioptas and Frank McSherry, *On spectral learning of mixtures of distributions*, Conference on Learning Theory (COLT) 2005, 458-469.

[2] Noga Alon and Nabil Kahale, *A spectral technique for coloring random 3-colorable graphs*, SIAM Journal on Computing **26** (1997), n. 6. 1733-1748.

[3] Noga Alon, Michael Krivelevich and Benny Sudakov, *Finding a large hidden clique in a random graph*, Proceedings of the 9^th Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.

[4] Sanjeev Arora and Ravi Kannan, *Learning mixtures of arbitrary gaussians*, Proceedings of the 32^{nd} annual ACM Symposium on Theory of computing (2001), 247-257.

[5] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry and Jared Saia, *Spectral analysis of data*, Proceedings of the 32^{nd} annual ACM Symposium on Theory of computing (2001), 619-626.

[6] Rajendra Bhatia, *Matrix Analysis*, New York, Springer-Verlag, 1997.

[7] Ravi Boppana, *Eigenvalues and graph bisection: an average case analysis*, Proceedings of the 28^{th} IEEE Symposium on Foundations of Computer Science (1987).

[8] Anirban Dasgupta, John Hopcroft and Frank McSherry, *Spectral analysis of random Graphs with skewed degree distributions*, Proceedings of the 42^{nd} IEEE Symposium on Foundations of Computer Science (2004), 602-610.

[9] Sanjoy Dasgupta and Leonard Schulman, *A two-round variant of EM for gaussian mixtures*, UAI (2000), 152-159.

[10] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan, *Competitive recommendation systems*, Proceedings of the 34^{th} ACM Symposium on Theory of Computing (STOC), pp. 82-90, 2002.

[11] Martin Dyer and Alan Frieze, *The solution of some random NP-hard problems in polynomial expected time*, Journal of Algorithms, **10**, 1989, 451-489.

[12] Uriel Feige and Joe Kilian, *Heuristics for semirandom graph problems*, Journal of Computer and System Sciences, **63**, 2001, 639-671.

[13] Uriel Feige and Eran Ofek, *Spectral techniques applied to sparse random graphs*, Random Structures and Algorithms, **27(2)**, 251–275, September 2005.

[14] Joel Friedman, Jeff Kahn and Endre Szemeredi, *On the second eigenvalue of random regular graphs*, Proceedings of the 21^{st} annual ACM Symposium on Theory of computing (1989), 587 - 598.

[15] Zoltan Furedi and Janos Komlos, *The eigenvalues of random symmetric matrices*, Combinatorica 1, **3**, (1981), 233–241.

[16] G. Golub, C. Van Loan (1996), *Matrix computations, third edition*, *The Johns Hopkins University Press Ltd., London.*

[17] Ravi Kannan, Hadi Salmasian and Santosh Vempala, *The spectral method for general mixture models*, Conference on Learning Theory (COLT) (2005), 444-457.

[18] L. Kucera, *Expected complexity of graph partitioning problems*, Discrete Applied Mathematics **57** (1995), 193-212.

[19] Frank McSherry, *Spectral partitioning of random graphs*, Proceedings of the 42^{nd} IEEE Symposium on Foundations of Computer Science (2001), 529-537.

[20] Christos Papadimitriou, Prabhakar Raghavan Hisao Tamaki and Santosh Vempala, *Latent semantic indexing: A probabilistic analysis*, Journal of Computer and System Sciences (special issue for PODS '01), **61**, (2000), 217-235.

[21] Mark Rudelson, *Random vectors in isotropic positions*, Journal of Functional Analysis, **164**, (1999), 60-72.

[22] Eugene Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Annals of Mathematics, **62**, (1955), 548-564.

[23] Eugene Wigner, *On the distribution of the roots of certain symmetric matrices*, Annals of Mathematics, **67**, (1958), 325-328.

[24] Santosh Vempala and Grant Wang, *A spectral algorithm for learning mixture models*, Journal of Computer and System Sciences, **68(4)**,(2004), 841-860.

[25] Van Vu, *Spectral norm of random matrices*, Proceedings of the 36^{th} annual ACM Symposium on Theory of computing (2005), 619-626.

Appendix

Linear Algebra Facts

We will use the following facts from linear algebra.

Fact 10. *For a matrix X with rank k, we have that $\|X\|_F^2 \leq k\|X\|_2^2$.*

Fact 11. ***(McSherry).*** *For a random matrix A, with $\bar{A} = \mathbf{E}[A]$, such that $\bar{A}$ has rank-k, we have that*

$$\|A^{(k)} - \bar{A}\|_F^2 \leq 8k\|A - \bar{A}\|_2^2.$$

The proof is by simple manipulation. For details, see [19].

Fact 12. *Since $\|X\|_F^2 = \sum_i \|X_i\|_2^2$, we have that the number of columns i such that $\|X_i\|^2$ is greater than $\|X\|_F^2/c$ is at most c.*

Fact 13. *If v and u be two vectors, and π_u and π_v be the projections onto these vectors, then as long as the difference $u - v$ is small with respect to the norm of v, the difference between the projections can be effectively bounded as a function of the difference between the vectors.*

$$\|\pi_u - \pi_v\|_2 \leq \frac{\|u - v\|_2}{\|v\|_2 - \|u - v\|_2} \tag{5}$$

The above fact is nothing but a very special case of Stewart's Theorem [16].

Proof of Lemma 2

Proof. With probability $1 - \frac{1}{poly(m,n)} - \frac{1}{n^{\log\theta - 8}}$,

$$\begin{aligned} \|A^{(k)} - \bar{A}\|_F^2 &\leq 8k\|A - \bar{A}\|^2 \\ &\leq 8k(\theta\sigma(4\sqrt{m} + 100\sqrt{\log n \log m}\sqrt{n})^2 \\ &\leq 8k(\theta\sigma(4\sqrt{m} + 100\sqrt{n \log n \log m})^2 \end{aligned}$$

Note that the first inequality is Fact 11, and the second is Corollary 7. □

Proof of Lemma 3

Proof. Using Fact 12, it is easy to see that if the result from Lemma 2 is true, then the number of columns i such that $\|A_i^{(k)} - \mathbf{E}[A_i]\|$ is greater than τ is at most $\frac{\|A^{(k)} - \mathbf{E}[A]\|_F^2}{\tau^2} \leq \frac{w_{\min}}{4k\log k}$. We call a sample "good" if it obeys $\|A_i^{(k)} - \mathbf{E}[A_i]\| \leq 2\tau$. Else, it is referred to as "bad". It is easy to see that if a good sample is picked as center in step (2a), then all good samples from the corresponding cluster are marked in the next step (2b) and are not picked henceforth. No good columns from any other cluster are picked. Thus, if we show that we only pick good samples in step (2a), we will be done, as there will then be exactly k columns picked, one from each cluster, and each of them will be close to its center.

The probability that the first column picked is good is given by $w_{\min}/4k\log k$. After picking p such columns, the total number of columns left is at least $(k-p+1)w_{\min}m$ and thus the probability of picking a good column at the i^{th} step is at most $\frac{1}{4(k-p+1)k\log k}$. Taking union bound over all the steps, the total probability of choosing a bad node is at most

$$\frac{1}{k\log k}\left(\frac{1}{4k}+\ldots+\frac{1}{4(k-p+1)}+\ldots+\frac{1}{8}\right)\leq\frac{\log k}{4k\log k}\leq\frac{1}{4k}$$

Thus, with probability $1-\frac{1}{4k}$, we have that all samples picked in step (2a) are good, and hence we have the claim in our lemma.

□

Proof of Lemma 8

Proof.

$$E_A\mathrm{Tr}\left(\left(AA^T-D\right)^l\right)=E_A\mathrm{Tr}\left(\left(AA^T-\int_X q(X)X\right)^l\right)$$

$$\leq E_AE_B\mathrm{Tr}\left(\left(AA^T-BB^T\right)^l\right)=E_AE_B\mathrm{Tr}\left(\left(\sum_{i=1}^m(A_iA_i^T-B_iB_i^T)\right)^l\right)$$

(Proposition (1)

$$=E_AE_BE_\zeta\mathrm{Tr}\left(\left(\sum_{i=1}^m\zeta_i(A_iA_i^T-B_iB_i^T)\right)^l\right)$$

(since $A_iA_I^T-B_iB_i^T$ is a symmetric random variable)

$$\leq 2^lE_AE_BE_\zeta\mathrm{Tr}\left(\left(\sum_{i=1}^m\zeta_iA_iA_i^T\right)^l\right)+lE_AE_BE_\zeta\mathrm{Tr}\left(\left(\sum_{i=1}^m\zeta_iB_iB_i^T\right)^l\right)$$

(Proposition (1))

$$=2^{l+1}E_AE_\zeta\mathrm{Tr}\left(\left(\sum_i\zeta_iA_iA_i^T\right)^l\right).$$

□