Discrete Applied Mathematics 203 (2016) 26–34

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Relating multiway discrepancy and singular values of nonnegative rectangular matrices

Marianna Bolla

Institute of Mathematics, Budapest University of Technology and Economics, Hungary

ARTICLE INFO

Article history:
Received 27 August 2014
Received in revised form 23 June 2015
Accepted 20 September 2015
Available online 23 October 2015

Keywords:
Discrepancy
Normalized matrix
Singular values
Spectral clusters

ABSTRACT

The minimum k-way discrepancy $\mathrm{md}_k(\mathbf{C})$ of a rectangular matrix $\mathbf{C}$ of nonnegative entries is the minimum of the maxima of the within- and between-cluster discrepancies that can be obtained by simultaneous k-clusterings (proper partitions) of its rows and columns. In Theorem 2, irrespective of the size of $\mathbf{C}$, we give the following estimate for the kth largest nontrivial singular value of the normalized matrix: $s_k \le 9\mathrm{md}_k(\mathbf{C})(k+2-9k\ln \mathrm{md}_k(\mathbf{C}))$, provided $0 < \mathrm{md}_k(\mathbf{C}) < 1$ and $k < \mathrm{rank}(\mathbf{C})$. This statement is a certain converse of Theorem 7 of Bolla (2014), and the proof uses some lemmas and ideas of Butler (2006), where the $k = 1$ case is treated. The result naturally extends to the singular values of the normalized adjacency matrix of a weighted undirected or directed graph.

1. Introduction

In many applications, for example when microarrays are analyzed, our data are collected in the form of an $m \times n$ rectangular matrix $\mathbf{C} = (c_{ij})$ of nonnegative real entries. We assume that $\mathbf{C}$ is *non-decomposable* (see Definition A.3.28 of [6]), i.e., $\mathbf{C}\mathbf{C}^T$ (when $m \le n$) or $\mathbf{C}^T\mathbf{C}$ (when $m > n$) is *irreducible*. Consequently, the row-sums $d_{row,i} = \sum_{j=1}^{n} c_{ij}$ and column-sums $d_{col,j} = \sum_{i=1}^{m} c_{ij}$ of $\mathbf{C}$ are strictly positive, and the diagonal matrices $\mathbf{D}_{row} = \mathrm{diag}(d_{row,1}, \dots, d_{row,m})$ and $\mathbf{D}_{col} = \mathrm{diag}(d_{col,1}, \dots, d_{col,n})$ are regular. Without loss of generality, we also assume that $\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} = 1$, since neither our main object, the *normalized matrix*

$$\mathbf{C}_D = \mathbf{D}_{row}^{-1/2}\mathbf{C}\mathbf{D}_{col}^{-1/2}, \tag{1}$$

nor the *multiway discrepancies* to be introduced are affected by the scaling of the entries of $\mathbf{C}$. It is known that the singular values of $\mathbf{C}_D$ are in the $[0, 1]$ interval. The positive ones, enumerated in non-increasing order, are the real numbers

$$1 = s_0 > s_1 \ge \cdots \ge s_{r-1} > 0,$$

where $r = \mathrm{rank}(\mathbf{C}_D) = \mathrm{rank}(\mathbf{C})$. Provided $\mathbf{C}$ is non-decomposable, 1 is a single singular value; it will be called *trivial* and denoted by s_0, since it corresponds to the trivial singular vector pair, which are disregarded in the clustering problems. This is a well-known fact of *correspondence analysis*, for further explanation see [6,7] and Section 3.

In Theorem 2, we will estimate the kth nontrivial singular value s_k of $\mathbf{C}_D$ from above with a (near zero, increasing) function of the *minimum k-way discrepancy* of $\mathbf{C}$ defined herein.

E-mail address: marib@math.bme.hu.

http://dx.doi.org/10.1016/j.dam.2015.09.013
0166-218X/

Definition 1. The multiway discrepancy of the rectangular matrix $\mathbf{C}$ of nonnegative entries in the proper k-partition $R_1, \dots, R_k$ of its rows and $C_1, \dots, C_k$ of its columns is

$$\mathrm{md}(\mathbf{C}; R_1, \dots, R_k, C_1, \dots, C_k) = \max_{\substack{1\le a,b\le k \\ X\subset R_a,\, Y\subset C_b}} \frac{|c(X,Y) - \rho(R_a, C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)|}{\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)}}, \tag{2}$$

where $c(X,Y) = \sum_{i\in X}\sum_{j\in Y} c_{ij}$ is the cut between $X \subset R_a$ and $Y \subset C_b$, $\mathrm{Vol}(X) = \sum_{i\in X} d_{row,i}$ is the volume of the row-subset X, $\mathrm{Vol}(Y) = \sum_{j\in Y} d_{col,j}$ is the volume of the column-subset Y, whereas $\rho(R_a, C_b) = \frac{c(R_a,C_b)}{\mathrm{Vol}(R_a)\mathrm{Vol}(C_b)}$ denotes the relative density between R_a and C_b. The minimum k-way discrepancy of $\mathbf{C}$ itself is

$$\mathrm{md}_k(\mathbf{C}) = \min_{\substack{R_1,\dots,R_k \\ C_1,\dots,C_k}} \mathrm{md}(\mathbf{C}; R_1, \dots, R_k, C_1, \dots, C_k).$$

In Section 3, we will extend this notion to an edge-weighted graph G and denote it by $\mathrm{md}_k(G)$. In that setup, $\mathbf{C}$ plays the role of the weighted adjacency matrix (symmetric in the undirected; quadratic, but usually not symmetric in the directed case), when the eigenvalues of the normalized adjacency matrix enter into the estimates, in their decreasing absolute values.

Note that $\mathrm{md}(\mathbf{C}; R_1, \dots, R_k, C_1, \dots, C_k)$ of (2) is the smallest α such that for every R_a, C_b pair and for every $X \subset R_a$, $Y \subset C_b$,

$$|c(X,Y) - \rho(R_a, C_b)\mathrm{Vol}(X)\mathrm{Vol}(Y)| \le \alpha\sqrt{\mathrm{Vol}(X)\mathrm{Vol}(Y)} \tag{3}$$

holds. Consequently, in the k-partitions of the rows and columns, giving the minimum k-way discrepancy (say, α^*) of $\mathbf{C}$, every R_a, C_b pair is α^*-*regular* in terms of the volumes, and α^* is the smallest possible discrepancy that can be attained with proper k-partitions. In the graph case, it resembles the notion of ϵ-regular pairs in the Szemerédi regularity lemma [18], albeit with given number of vertex-clusters, which are usually not equitable; further, with volumes, instead of cardinalities.

Though it is not always called discrepancy, this notion has intensively been used since the 1970s, e.g., in [9] and [18–20]. Thomason [19,20] introduced it in the context of what he called (p, α)-jumbled graphs and proved relations between this and similar notions, related to pseudo-random graphs. Expander graphs and the expander mixing lemma for simple regular graphs are also closely related to this notion, e.g., Alon, Spencer, Hoory, Linial, Widgerson [3,15]. Bollobás and Nikiforov [10] extended the notion of discrepancy to Hermitian matrices. Then they defined two types of discrepancy for graphs and showed that their estimate is valid to both, with due regard to a theorem of Thomason [20]. They also proved that for a large graph G, one type of these discrepancies closely approximates the discrepancy of its adjacency matrix (as a real Hermitian matrix). In Chung, Graham, Wilson [13], the authors used the term quasi-random for simple graphs that satisfy any of some equivalent properties, some of which closely related to discrepancy and eigenvalue separation.

Here we rather extend the notion of discrepancy used by Chung and Graham for simple graphs with given degree sequences. In [12], the authors proved that for simple graphs 'small' discrepancy $\mathrm{disc}(G)$ (with our notation, $\mathrm{md}_1(G)$) is caused by eigenvalue 'separation': the second largest singular value (which is also the second largest absolute value eigenvalue), s_1, of the normalized adjacency matrix is 'small', i.e., separated from the trivial singular value $s_0 = 1$, which is the edge of the spectrum. More exactly, they proved $\mathrm{disc}(G) \le s_1$, hence giving some kind of generalization of the *expander mixing lemma for irregular graphs*.

In the backward direction, Bollobás and Nikiforov [10] estimated the second largest singular value of an $n \times n$ Hermitian matrix $\mathbf{A}$ by $C\mathrm{disc}(\mathbf{A}) \log n$, and showed that this is best possible up to a multiplicative constant. Bilu and Linial [4] proved the converse of the expander mixing lemma for simple regular graphs, but their key lemma, producing this statement, goes beyond regular graphs, see Section 3.1 for details. In Alon et al. [2], the authors relaxed the notion of eigenvalue separation to essential eigenvalue separation (by introducing a parameter for it, and requiring the separation only for the eigenvalues of a relatively large part of the graph). Then they proved relations between the constants of this kind of eigenvalue separation and discrepancy.

For a general rectangular matrix $\mathbf{C}$ of nonnegative entries, Butler [11] proved the following forward and backward statement in the $k = 1$ case:

$$\mathrm{disc}(\mathbf{C}) \le s_1 \le 150\mathrm{disc}(\mathbf{C})(1 - 8\ln \mathrm{disc}(\mathbf{C})), \tag{4}$$

where $\mathrm{disc}(\mathbf{C})$ is our $\mathrm{md}_1(\mathbf{C})$ and, with our notation, s_1 is the largest nontrivial singular value of $\mathbf{C}_D$ (he denoted it with σ_2). Since $s_1 < 1$, the upper estimate makes sense for very small discrepancy, in particular, for $\mathrm{disc}(\mathbf{C}) \le 8.868 \times 10^{-5}$. The lower estimate further generalizes the expander mixing lemma to rectangular matrices.

So far, the overall discrepancy has been considered in the sense, that $\mathrm{disc}(\mathbf{C})$ or $\mathrm{disc}(G)$ measures the largest possible deviation between the actual and expected connectedness of arbitrary (sometimes disjoint) subsets X, Y, where under expected the hypothesis of independence is understood (which corresponds to the rank 1 approximation of the underlying matrix). Our purpose is, in the multicluster scenario, to find similar relations between the minimum k-way discrepancy and the SVD of the normalized matrix, for given k. In the *backward direction*, in Section 2, we will prove the following.