Discrete Applied Mathematics 176 (2014) 3–11

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

SVD, discrepancy, and regular structure of contingency tables

Marianna Bolla *

Institute of Mathematics, Budapest University of Technology and Economics, Hungary
Center for Telecommunication and Informatics, Debrecen University, Hungary

ARTICLE INFO

Article history:
Received 15 November 2012
Received in revised form 25 February 2014
Accepted 30 March 2014
Available online 18 April 2014

Keywords:
Normalized contingency table
Normalized two-way cuts
Biclustering
Discrepancy
Directed graphs

ABSTRACT

Factors, obtained by correspondence analysis, are used to find biclustering of a contingency table such that the row–column cluster pairs are regular, i.e., they have small discrepancy. In our main theorem, the constant of the so-called volume-regularity is related to the SVD of the normalized contingency table. This result is applicable to two-way cuts when both the rows and columns are divided into the same number of clusters, thus extending partly the result of Butler for estimating the discrepancy of a contingency table by the largest non-trivial singular value of the normalized table (one-cluster, rectangular case), and partly the result of Bolla for estimating the constant of volume-regularity by the structural eigenvalues and the distances of the corresponding eigen-subspaces of the normalized modularity matrix of an edge-weighted graph (several clusters, symmetric case).

1. Introduction

A typical problem of contemporary cluster analysis is to find relatively small number of groups of objects, belonging to rows and columns of a contingency table which exhibit homogeneous behavior with respect to each other and do not differ significantly in size. To make inferences on the separation that can be achieved for a given number of clusters, minimum normalized two-way cuts and discrepancies of the cluster pairs are investigated and related to the SVD of the normalized contingency table.

Contingency tables are rectangular arrays with nonnegative, real entries, e.g., the keyword-document matrix or microarray gene expression data. In the former one, the matrix entries are associations between documents and words, whereas in the latter one, they are expression levels of genes under different conditions. We also look for a bipartition of the genes and conditions such that genes in the same cluster equally influence conditions of the same cluster. To find so-called biclustering, i.e., simultaneous clustering of the rows and columns, a great variety of algorithms are used.

The first algorithm of this flavor is due to Hartigan [17], who used two-way analysis of variance techniques to find constant valued submatrices within the rectangular array. In [24], applications to microarrays is presented, where biclusters identify subsets of genes sharing similar expression patterns across subsets of conditions, but the authors do not use spectral methods. We will rather concentrate on methods that use the SVD of the original or normalized contingency table.

The Latent Semantic Indexing offers an SVD-based algorithm, which can be generalized in many different ways. For example, in [15], the authors find scoring systems simultaneously for the keywords and documents with respect to the most important topics or factors, and use the singular vector pairs corresponding to the k outstanding singular values of the table.

If a scoring system is endowed with the marginal measures, the problem can be formulated in terms of the correspondence analysis, based on the SVD of the normalized table, see [4,13,16]. We will show, how in possession of the

* Correspondence to: Institute of Mathematics, Budapest University of Technology and Economics, Hungary. Tel.: +36 1 200 0646; fax: +36 1 4631677.
E-mail addresses: marib@math.bme.hu, marianna.bolla@gmail.com.

http://dx.doi.org/10.1016/j.dam.2014.03.022
0166-218X/

correspondence factor pairs a biclustering can be performed that finds simultaneous clustering of the rows and columns of the table such that certain regularity requirements are met.

A survey of biclustering algorithms in data mining, especially in biological data analysis is given in [8,24]. To find biclustering of a binary table via the k-means algorithm is also discussed in [10], where the author embeds the contingency table into a bipartite graph and uses normalized cut objectives and SVD to obtain the convenient biclustering. To find the SVD for large rectangular matrices, randomized algorithms are favored. A randomized, so-called fast Monte Carlo algorithm for the SVD and its application for clustering large graphs via the k-means algorithm is presented in [12].

The problem is also related to the Page-rank (see [19]). As for the microarray analysis, the authors of [20] use the leading singular values and vector pairs of the normalized contingency table to find a so-called checkerboard pattern in it, but they do not give estimation how this pattern approaches the original table. Some authors, e.g., [21], impose sparsity inducing conditions on the leading singular vector pairs, so that they have piecewise constant structure with many zero coordinates, and so, produce a checkerboard structure.

Though, many papers deal with the SVD-based biclustering of the underlying contingency table (see also [18,22,26]), they just introduce numerical algorithms, possibly with some constraints, which utilize the well-known favorable properties of low-rank approximations. After finding the checkerboard patterns, no inference is made on the homogeneity of the so obtained biclusters by means of the SVD of the table. It is also a drawback that the low-rank approximation, unlike the original table, may have negative entries.

In Section 2, we relate the biclustering problem to normalized two-way cuts of contingency tables, akin to the way normalized cuts of edge-weighted graphs are estimated by the normalized Laplacian spectra, see [10]. The minimization of this objective function favors biclusterings with dense diagonal, and sparse off-diagonal blocks. In terms of microarrays, it finds partition of genes and conditions into the same number of clusters such that to each cluster of conditions we can find a collection of genes responsible for this condition, and vice versa.

In Section 3, more generally, we are looking for so-called volume-regular row–column clusters pairs, such that the association between their row and column subsets is homogeneous, but not necessarily dense or sparse. The minimum of the pairwise discrepancies is related to the so-called structural singular values and corresponding eigen-functions of the normalized table. We use the one-cluster estimation of Butler [9], who estimates the discrepancy of the whole contingency table by the largest non-trivial singular value of the normalized table (one-cluster, rectangular case); moreover, he proves a two-sided relation between this singular value and the discrepancy. Here we extend the forward direction of this estimation for the k-cluster case, where our results also indicate the optimal choice of k. For this purpose, we use the result of Bolla [3] for estimating the constant of volume-regularity by the structural eigenvalues and the distances of the corresponding eigen-subspaces of the normalized modularity matrix of an edge-weighted graph. Since there are several eigenvalues (singular values) responsible for this versatile property, together with the corresponding eigenvectors (singular vector pairs), this statement is more complicated to prove and cannot be simply inverted, akin to the one-cluster case. Nonetheless, this problem has not yet been treated in the literature.

Section 4 is devoted to discussion and possible extension to directed graphs.

2. Normalized two-way cuts of contingency tables

Let $\mathbf{C}$ be a contingency table on row set $Row = \{1, \dots, n\}$ and column set $Col = \{1, \dots, m\}$, where $\mathbf{C}$ is $n \times m$ matrix of entries $c_{ij} \geq 0$. Without loss of generality, we suppose that there are no identically zero rows or columns. Here c_{ij} is some kind of association between the objects behind row i and column j, where 0 means no interaction at all.

Let the row- and column-sums of $\mathbf{C}$ be

$$d_{row,i} = \sum_{j=1}^{m} c_{ij} \quad (i = 1, \dots, n) \quad \text{and} \quad d_{col,j} = \sum_{i=1}^{n} c_{ij} \quad (j = 1, \dots, m)$$

which are collected in the main diagonals of the $n \times n$ and $m \times m$ diagonal matrices $\mathbf{D}_{row}$ and $\mathbf{D}_{col}$, respectively. The matrix

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

is called the *correspondence matrix (normalized contingency table)* belonging to the table $\mathbf{C}$, see [4]. If we multiply all the entries of $\mathbf{C}$ with the same positive constant, the correspondence matrix $\mathbf{C}_{corr}$ will not change. Therefore, without the loss of generality, $\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} = 1$ will be assumed in the sequel.

Given an integer $k (0 < k \leq \mathtt{rank}(\mathbf{C}))$, we want to simultaneously partition the rows and columns of $\mathbf{C}$ into disjoint, nonempty subsets

$$Row = R_1 \cup \dots \cup R_k, \qquad Col = C_1 \cup \dots \cup C_k$$

such that we impose conditions on the *cuts* $c(R_a, C_b) = \sum_{i \in R_a} \sum_{j \in C_b} c_{ij} (a, b = 1, \dots, k)$ between the row–column cluster pairs. For this purpose, the following so-called *normalized two-way cut* of the contingency table with respect to the above k-partitions $P_{row} = (R_1, \dots, R_k)$ and $P_{col} = (C_1, \dots, C_k)$ of its rows and columns and the collection of signs σ is defined as