SVD, discrepancy, and regular structure of contingency tables

Marianna Bolla
Institute of Mathematics
Technical University of Budapest
(Research supported by the TÁMOP-4.2.2.B-10/1-2010-0009 project)
marib@math.bme.hu
European Meeting of Statisticians, Budapest

July 21, 2013

Outline

- **Motivation**:
  - 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 bicuts are investigated and related to the SVD of the correspondence matrix.
- **Topics**:
  - singular value decomposition (SVD) of a correspondence matrix;
  - SVD and normalized bicuts of the contingency table;
  - volume-regular row–column clusters pairs;
  - application and possible extension to directed graphs.

References

- We partly extend the result of
  Butler, S., Using discrepancy to control singular values for nonnegative matrices, *Lin. Alg. Appl.* (2006)
  for estimating the discrepancy of a contingency table by the second largest singular value of the normalized table (one-cluster, rectangular case),
- and the result of
  B, M., Modularity spectra, eigen-subspaces, and structure of weighted graphs, *European Journal of Combinatorics* (2013)
  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.

SVD of contingency tables and correspondence matrices

$\mathbf{C} = (c_{ij})$: $n \times m$ contingency table, $c_{ij} \geq 0$.
Row set: $Row = \{1, \ldots, n\}$
Column set: $Col = \{1, \ldots, m\}$

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

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

$$\mathbf{D}_{row} = \mathtt{diag}\,(d_{row,1}, \ldots, d_{row,n}) \quad \mathbf{D}_{col} = \mathtt{diag}\,(d_{col,1}, \ldots, d_{col,m}).$$

Representation

For a given integer $1 \leq k \leq \min\{n, m\}$, we are looking for k-dimensional representatives $\mathbf{r}_1, \ldots, \mathbf{r}_n$ of the rows and $\mathbf{c}_1, \ldots, \mathbf{c}_m$ of the columns such that they minimize the objective function

$$Q_k = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \|\mathbf{r}_i - \mathbf{c}_j\|^2 \tag{1}$$

subject to

$$\sum_{i=1}^{n} d_{row,i} \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_k, \quad \sum_{j=1}^{m} d_{col,j} \mathbf{c}_j \mathbf{c}_j^T = \mathbf{I}_k. \tag{2}$$

When minimized, the objective function Q_k favors k-dimensional placement of the rows and columns such that representatives of highly associated rows and columns are forced to be close to each other. As we will see, this is equivalent to the problem of correspondence analysis.

Solution

$$\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_k) = (\mathbf{r}_1^T, \ldots, \mathbf{r}_n^T)^T \; n \times k$$
$$\mathbf{Y} := (\mathbf{y}_1, \ldots, \mathbf{y}_k) = (\mathbf{c}_1^T, \ldots, \mathbf{c}_m^T)^T \; m \times k$$

$$Q_k = 2k - \mathtt{tr}\,(\mathbf{D}_{row}^{1/2}\mathbf{X})^T \mathbf{C}_{corr} (\mathbf{D}_{col}^{1/2}\mathbf{Y}) \to \min$$

subject to

$$\mathbf{X}^T \mathbf{D}_{row} \mathbf{X} = \mathbf{I}_k, \quad \mathbf{Y}^T \mathbf{D}_{col} \mathbf{Y} = \mathbf{I}_k,$$

where $\mathbf{C}_{corr} = \mathbf{D}_{row}^{-1/2} \mathbf{C} \mathbf{D}_{col}^{-1/2}$: correspondence matrix (normalized contingency table) belonging to the table $\mathbf{C}$.

Representation theorem

Let $\mathbf{C}_{corr} = \sum_{i=1}^{r} s_i \mathbf{v}_i \mathbf{u}_i^T$ be SVD, where $r \leq \min\{n, m\}$ is the rank of $\mathbf{C}_{corr}$, or equivalently (since there are not identically zero rows or columns), the rank of $\mathbf{C}$ and $1 = s_1 \geq s_2 \geq \cdots \geq s_r > 0$.
$\mathbf{v}_1 = (\sqrt{d_{row,1}}, \ldots, \sqrt{d_{row,n}})^T$ and $\mathbf{u}_1 = (\sqrt{d_{col,1}}, \ldots, \sqrt{d_{col,m}})^T$.
Let $k \leq r$ be a positive integer such that $s_k > s_{k+1}$. Then

$$\min Q_k = 2k - \sum_{i=1}^{k} s_i$$

and it is attained with $\mathbf{X}^* = \mathbf{D}_{row}^{-1/2}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ and $\mathbf{Y}^* = \mathbf{D}_{col}^{-1/2}(\mathbf{u}_1, \ldots, \mathbf{u}_k)$.

Normalized bicuts of contingency tables

Given an integer k $(0 < k \leq r)$, we want to simultaneously partition the rows and columns into disjoint, nonempty subsets

$$Row = R_1 \cup \cdots \cup R_k, \quad Col = C_1 \cup \cdots \cup C_k$$

such that the cuts $c(R_a, C_b) = \sum_{i \in R_a} \sum_{j \in C_b} c_{ij}$ $(a, b = 1, \ldots, k)$ between the row-column cluster pairs be as **homogeneous** as possible.

The normalized two-way cut of $\mathbf{C}$ given the above k-partitions $P_{row} = (R_1, \ldots, R_k)$ and $P_{col} = (C_1, \ldots, C_k)$ and the collection of signs σ:

$$\nu_k(P_{row}, P_{col}, \sigma) =$$

$$\sum_{a=1}^{k} \sum_{b=1}^{k} \left(\frac{1}{\mathtt{Vol}(R_a)} + \frac{1}{\mathtt{Vol}(C_b)} + \frac{2\sigma_{ab}\delta_{ab}}{\sqrt{\mathtt{Vol}(R_a)\mathtt{Vol}(C_b)}} \right) c(R_a, C_b),$$

where

$$\texttt{Vol}(R_a) = \sum_{i \in R_a} d_{row,i}, \quad \texttt{Vol}(C_b) = \sum_{j \in C_b} d_{col,j},$$

δ_{ab} is the Kronecker delta, $\sigma_{ab} = \pm 1$, and $\sigma = (\sigma_{11}, \ldots, \sigma_{kk})$.
For the normalized bicut of $\mathbf{C}$:

$$\min_{P_{row}, P_{col}, \sigma} \nu_k(P_{row}, P_{col}, \sigma) \geq 2k - \sum_{i=1}^{k} s_i.$$

Special case: edge-weighted graph

$\mathbf{W}$: symmetric $n \times n$ edge-weigh matrix.
$\mathbf{D} = \mathbf{D}_{row} = \mathbf{D}_{col}$, $\mathbf{W}_{corr} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$: normalized modularity matrix, $\mathbf{I} - \mathbf{W}_{corr}$: normalized Laplacian.

- When the $k-1$ largest absolute value eigenvalues of $\mathbf{W}_{corr}$ are all positive: $\mathbf{r}_i = \mathbf{c}_i$ $(i = 1, \ldots, n = m)$. With the choice $\sigma_{bb} = 1$ $(b = 1, \ldots, k)$, $\nu_k(P_{row}, P_{col}, \sigma)$ is twice the normalized cut of the graph with respect to the k-partition $P_{row} = P_{col}$ of the vertices. The normalized bicut favors k-partitions with low inter-cluster edge-densities: **community structure**.
- When the $k-1$ largest absolute value eigenvalues of the normalized modularity matrix are all negative: $\mathbf{r}_i = -\mathbf{c}_i$ and the minimum of the normalized bicut is attained with the choice $\sigma_{bb} = -1$ $(b = 1, \ldots, k)$. The normalized bicut favors k-partitions with low intra-cluster edge-densities: **anticommunity structure**.

Regular row-column cluster pairs

The Expander Mixing Lemma for edge-weighted graphs naturally extends to this situation (Butler): for all $R \subset Row$ and $C \subset Col$

$$|c(R, C) - \mathtt{Vol}(R)\mathtt{Vol}(C)| \leq s_2 \sqrt{\mathtt{Vol}(R)\mathtt{Vol}(C)},$$

where s_2 is the largest but 1 singular value of $\mathbf{C}_{corr}$.
Since the spectral gap of $\mathbf{C}_{corr}$ is $1 - s_2$, in view of the above Expander Mixing Lemma, 'large' spectral gap is an indication of 'small' discrepancy: the weighted cut between any row and column subset of the contingency table is near to what is expected in a random table.

Volume-regular cluster pairs

We extend the notion of discrepancy to volume-regular pairs.

Definition

The row–column cluster pair $R \subset Row$, $C \subset Col$ of the contingency table $\mathbf{C}$ of total volume 1 is γ-volume regular if for all $X \subset R$ and $Y \subset C$ the relation

$$|c(X,Y) - \rho(R,C)\mathtt{Vol}(X)\mathtt{Vol}(Y)| \leq \gamma\sqrt{\mathtt{Vol}(R)\mathtt{Vol}(C)}$$

holds, where $\rho(R,C) = \frac{c(R,C)}{\mathtt{Vol}(R)\mathtt{Vol}(C)}$ is the relative inter-cluster density of the row–column pair R, C.

We will show that for given k, if the clusters are formed via applying the weighted k-means algorithm for the optimal row- and column representatives, respectively, then the so obtained row–column cluster pairs are homogeneous in the sense that they form equally dense parts of the contingency table.

Weighted k-variance

The weighted k-variance of the k-dimensional row representatives is defined by

$$S_k^2(\mathbf{X}) = \min_{(R_1,\ldots,R_k)} \sum_{a=1}^{k} \sum_{j \in R_a} d_{row,j} \|\mathbf{r}_j - \bar{\mathbf{r}}_a\|^2,$$

where $\bar{\mathbf{r}}_a = \frac{1}{\mathtt{Vol}(R_a)} \sum_{j \in R_a} d_{row,j} \mathbf{r}_j$ is the weighted center of cluster R_a $(a = 1, \ldots, k)$. Similarly, the weighted k-variance of the k-dimensional column representatives is

$$S_k^2(\mathbf{Y}) = \min_{(C_1,\ldots,C_k)} \sum_{a=1}^{k} \sum_{j \in C_a} d_{col,j} \|\mathbf{c}_j - \bar{\mathbf{c}}_a\|^2,$$

where $\bar{\mathbf{c}}_a = \frac{1}{\mathtt{Vol}(C_a)} \sum_{j \in C_a} d_{col,j} \mathbf{c}_j$ is the weighted center of cluster C_a $(a = 1, \ldots, k)$. Observe, that the trivial vector components can be omitted, and the k-variance of the so obtained $(k-1)$-dimensional representatives will be the same.

Theorem

Theorem

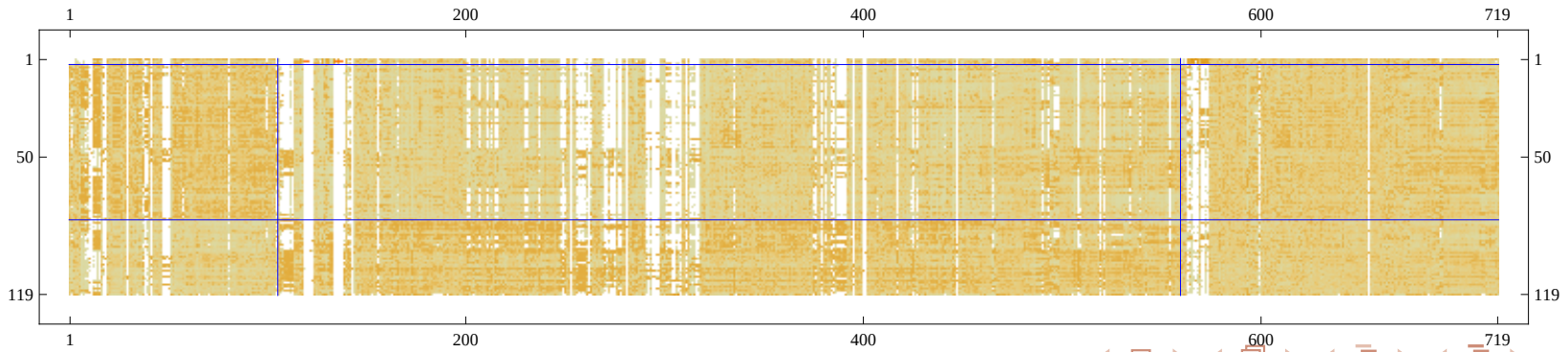
Let $\mathbf{C}$ *be a contingency table of n-element row set Row and m-element column set Col, with row- and column sums* $d_{row,1}, \ldots, d_{row,n}$ *and* $d_{col,1}, \ldots, d_{col,m}$, *respectively. Suppose that* $\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} = 1$ *and there are no dominant rows and columns:* $d_{row,i} = \Theta(1/n)$, $(i = 1, \ldots, n)$ *and* $d_{col,j} = \Theta(1/m)$, $(j = 1, \ldots, m)$ *as* $n, m \to \infty$. *Let the singular values of* $\mathbf{C}_{corr}$ *be*

$$1 = s_1 > s_2 \geq \cdots \geq s_k > \varepsilon \geq s_i, \quad i \geq k + 1.$$

The partition $(R_1, \ldots, R_k)$ *of Row and* $(C_1, \ldots, C_k)$ *of Col are defined so that they minimize the weighted k-variances* $S_k^2(\mathbf{X})$ *and* $S_k^2(\mathbf{Y})$ *of the row and column representatives. Suppose that there are constants* $0 < K_1, K_2 \leq \frac{1}{k}$ *such that* $|R_i| \geq K_1 n$ *and* $|C_i| \geq K_2 m$ $(i = 1, \ldots, k)$, *respectively. Then the* R_i, C_j *pairs are* $\mathcal{O}(\sqrt{2k}(S_k(\mathbf{X})S_k(\mathbf{Y})) + \varepsilon)$*-volume regular* $(i, j = 1, \ldots, k)$.

Application

We applied the biclustering algorithm to find simultaneously clusters of stores and products based on their consumption in TESCO stores. We found 3 clusters of the stores in which the consumption of the products belonging to the same cluster was homogeneous with consumption-density $\frac{c(R_a, C_b)}{\mathtt{Vol}(R_a)\mathtt{Vol}(C_b)}$ between store-cluster R_a and product-cluster C_b $(a, b = 1, \ldots, 3)$. After sorting the rows and columns according to their cluster memberships, we plotted the entries $\frac{c_{ij}}{d_{row,i}d_{col,j}}$ (there was one exceptional store-cluster which contained only 3 stores, but the others could be identified with groups of smaller and larger stores associated with product groups of high consumption-density).



Directed graphs

$\mathbf{W}$: $n \times n$ (not symmetric) weight matrix of a directed graph, where w_{ij} is the weight of the $i \to j$ edge $(i, j = 1, \ldots, n;\ i \neq j)$, and $w_{ii} = 0$ $(i = 1, \ldots, n)$.
The generalized in- and out-degrees:

$$d_{out,i} = \sum_{j=1}^{n} w_{ij} \quad (i = 1, \ldots, n)$$

$$d_{in,j} = \sum_{i=1}^{n} w_{ij} \quad (j = 1, \ldots, n).$$

$\mathbf{D}_{in} = \mathtt{diag}\,(d_{in,1}, \ldots, d_{in,n})$ and $\mathbf{D}_{out} = \mathtt{diag}\,(d_{out,1}, \ldots, d_{out,n})$ are the in- and out-degree matrices. Suppose that there are no sources and sinks (i.e. no zero out- and in-degrees).

$$\mathbf{W}_{corr} = \mathbf{D}_{out}^{-1/2} \mathbf{W} \mathbf{D}_{in}^{-1/2},$$

and its SVD is used to minimize the normalized bicut of $\mathbf{W}$ as a contingency table.

Regular in- and out-vertex cluster pairs

The V_{in}, V_{out} in- and out-vertex cluster pair of the directed graph (with sum of the weights of directed edges 1) is γ-volume regular if for all $X \subset V_{out}$ and $Y \subset V_{in}$ the relation

$$|w(X,Y) - \rho(V_{out}, V_{in})\mathtt{Vol}_{out}(X)\mathtt{Vol}_{in}(Y)| \leq \gamma\sqrt{\mathtt{Vol}_{out}(V_{out})\mathtt{Vol}_{in}(V_{in})}$$

holds, where the directed cut $w(X,Y)$ is the sum the weights of the $X \to Y$ edges,
$\mathtt{Vol}_{out}(X) = \sum_{i\in X} d_{out,i}$, $\mathtt{Vol}_{in}(Y) = \sum_{j\in Y} d_{in,j}$, and
$\rho(V_{out}, V_{in}) = \frac{w(V_{out},V_{in})}{\mathtt{Vol}_{out}(V_{out})\mathtt{Vol}_{in}(V_{in})}$ is the relative inter-cluster density of the out–in cluster pair V_{out}, V_{in}. The clustering $(V_{in,1}, \ldots, V_{in,k})$ and $(V_{out,1}, \ldots, V_{out,k})$ of the columns and rows – guaranteed by the above theorem – corresponds to in- and out-clusters of the same vertex set such that the directed information flow $V_{out,a} \to V_{in,b}$ is as homogeneous as possible for all $a, b = 1, \ldots, k$ pairs. **Emigration–immigration patterns of countries**.

Thank you for your attention