Graph Partitioning by Correspondence Analysis and Taxicab Correspondence Analysis

Vartan Choulakian Jules de Tibeiro

Université de Moncton
Moncton, NB, Canada

Université de Moncton
Shippagan, NB, Canada

March 5, 2014

Table of contents

Background

- Correspondence Analysis (CA) and Taxicab Correspondence Analysis (TCA) of relational datasets can mathematically be described as weighted loopless graphs.
- Such data appear in network analysis, see for instance, Kolaczyk (2009).
- TCA is a L_1 version of CA recently proposed by Choulakian (2006a), and will be briefly described later in the talk.
- For details about CA and PCA, interested readers can consult Greenacre (1984), Lebart (1984), Benzécri (1992), or Saporta (2011).
- Fraser and Hunter (1975) collected data on congenital cardiovascular defects, named as FH dataset, which is an incomplete paired categorical data.

Background

Table 1: Observed counts of pairs of cardiac malformations.

	ToF	*VSD*	PS	*TGV*	PDA	AS	ASD	Tru	TA	CoA	Dex	Ptr	A-V
ToF		13	19	10	4	1	1	0	1	0	1	2	0
VSD			3	5	3	3	6	1	0	0	2	1	0
PS				2	0	1	1	3	1	0	0	0	0
TGV					4	1	2	1	0	1	0	0	0
PDA						2	0	1	2	0	0	0	1
AS							2	0	1	3	2	0	0
ASD								0	1	1	0	0	1
Tru									0	0	0	1	0
TA										0	0	0	0
CoA											0	0	0
Dex												0	0
Ptr													0
A-V													

Motivation

- A congenital cardiovascular defect occurs when the heart or blood vessels near the heart do not develop normally before birth.
- Congenital cardiovascular defects are present in about 1 percent of live births.
- The goal of Fraser and Hunter (1975) was to reveal etiologic relations among cardiac lesions: Sibships in which two or more children had dissimilar cardiac lesions.
- The diagonal of the FH dataset (Table 1) is incomplete which means that cases where two siblings had the same defect were not recorded.

Motivation

- Additionally, the pairings are unordered, so only the top off-diagonal entries in the table are recorded, representing the total number of pairs, $n = 111$, occuring in either order.
- The FH dataset was first analysed by MacGibbon (1983).
- Dinwoodie and MacGibbon (2003) rejected the quasi-independence model, $p_{ij} = \alpha_i \alpha_j$ for $i \neq j$, of the heart malformations using exact inference procedure.
- de Tibeiro (1996) reconstructed the diagonal entries of Table 1 by minimizing the trace criterion and applied CA to the reconstructed dataset. He interpreted only the first principal axis, because the second principal axis was not direct. There was an inversion problem.

Motivation

- de Tibeiro and Murdoch (2010) applied CA after imputing the missing diagonal cells via a Bayesian procedure, however the first principal factor was inversed.
- Benzecri (1973) discussed extensively CA of such data sets highlighting three related major problems in applying CA: (i) the influence of the diagonal elements on the factors and dispersion measures, (ii) the interpretation of indirect factors, and (iii) the eigenvalues have very high values near 1 which does not imply a partition of the data into blocks.
- One of our major aim here is to alleviate these three stated problems by introducing a new coding of the FH dataset much used in graph theory, then applying CA and TCA to the coded dataset.

Introduction

- $\mathbf{Y} = (y_{ij})$ represent the symmetrized FH dataset, $y_{ij} = y_{ji} \geq 0$ for $i \neq j$ and $y_{ij} = 0$ for $i = 1, 2, \ldots, c$.
- G be a loopless undirected graph with vertex set $V = \{v_1, \ldots, v_{13}\}$ and multiple edge set $E = \{e_1, \ldots, e_{111}\}$.
- We consider $\mathbf{Y}$ as the adjacency matrix or the weight matrix of the graph G, where y_{ij} is the number of edges in G with endpoints $\{v_i, v_j\}$.
- y_{ij} measures the degree of association between the *ith* and *jth* column variables or vertices.

Introduction

- The edge-vertex incidence matrix $\mathbf{Z} = (z_{ij})$ of the graph G is the $n\ x\ c = 111\ x\ 13$ matrix in which entry z_{ij} is 1 if the endpoint $v_j \in e_i$ and otherwise 0.
- Each edge has 2 endpoints, because the graph is not directed. So each row of $\mathbf{Z}$ has only 2 ones and the other entries are 0.
- For example, consider the second nonzero entry, 19 (Table 1), there are 19 edges (pairs of siblings) with $\{v_1, v_3\} = \{ToF, PS\}$. Anyone of the l9 edges can be represented as a 0 or 1 edge-vertex incidence indicator vector of size 13,

$$\mathbf{(1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)} \qquad (1)$$

Introduction

- The vector in (1) is repeated 19 times in $\mathbf{Z}$.
- The edge-vertex incidence matrix $\mathbf{Z}$ can also be represented as a weighted incidence matrix, $\mathbf{W}$, of size 39×13, where 39 represents the number of nonzero entries in Table 1 and the second nonzero entry will be coded as,

$$\mathbf{(19 \quad 0 \quad 19 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)} \qquad (2)$$

- The matrices $\mathbf{Z}$ and $\mathbf{W}$ have no rows with zero counts. Let, $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$

Introduction

Table 2: The B Matrix.

	ToF	*VSD*	PS	*TGV*	PDA	AS	ASD	Tru	TA	CoA	Dex	Ptr	A-V
ToF	52	13	19	10	4	1	1	0	1	0	1	2	0
VSD	13	37	3	5	3	3	6	1	0	0	2	1	0
PS	19	3	30	2	0	1	1	3	1	0	0	0	0
TGV	10	5	2	26	4	1	2	1	0	1	0	0	0
PDA	4	3	0	4	17	2	0	1	2	0	0	0	1
AS	1	3	1	1	2	16	2	0	1	3	2	0	0
ASD	1	6	1	2	0	2	15	0	1	1	0	0	1
Tru	0	1	3	1	1	0	0	7	0	0	0	1	0
TA	1	0	1	0	2	1	1	0	6	0	0	0	0
CoA	0	0	0	1	0	3	1	0	0	5	0	0	0
Dex	1	2	0	0	0	2	0	0	0	0	5	0	0
Ptr	2	1	0	0	0	0	0	1	0	0	0	4	0
A-V	0	0	0	0	1	0	1	0	0	0	0	0	2

Introduction

- The degree of the vertex v_i, $degree(v_i) = \sum_{j=1}^{13} y_{ij}$ is defined as the number of edges incident to the vertex v_i and the diagonal matrix with diagonal elements $degree(v_i)$, $\mathbf{D} = \mathbf{Diag}(degree(v_i))$.
- We then have $\mathbf{B} = \mathbf{D} + \mathbf{Y}$ which represents the original data set $\mathbf{Y}$ with imputed or reconstructed diagonal elements (Benzecri (1973)).
- So CA can be applied to $\mathbf{B}$, as well as to $\mathbf{Z}$; traces of this fact can also be found in Seary and Richards (1995).

Introduction

- Finding group structures in a graph, i.e., the graph partitioning problem, is an important aim in Statistics.
- We present here CA with continuous relaxed version and TCA as a discrete relaxed version for the graph partitioning problem.

Problem Formulation

- We follow Ding (2004) or von Luxburg (2007) in considering the problem of partitioning V, the set of vertices, into two disjoint classes A and $\overline{A}$ with some specific optimal property.
- Let $\mathbf{u} = \{-1, +1\}^c$ be a cut, a membership indicator vector of size c, such that $u_i = 1$ if $v_i \in A$, and $u_i = -1$ if $v_i \in \overline{A}$.
- Our aim is to find good classes A and $\overline{A}$, such that the weight of the edges between A and $\overline{A}$ is as low as possible.

Problem Formulation

- We define the volume between A and $\overline{A}$ to be,

$$volBetween(A, \overline{A}) = \sum_{i \in A} \sum_{j \in \overline{A}} y_{ij} \tag{3}$$

  which is a dissimilarity index between the subgraphs defined by the subsets A and $\overline{A}$.

- Furthermore, we want the groups A and $\overline{A}$ to be balanced. So, the graph partitioning problem is stated as a combinatorial optimization problem defined by minimizing the loss function,

Problem Formulation

$$
\begin{aligned}
\min_{\mathbf{u}=\{-1,+1\}^c} \quad & Cutsize(\mathbf{u}) = 2volBetween(A, \overline{A}) \\
= \quad & \frac{1}{4}\sum_{i=1}^{c}\sum_{j=1}^{c} y_{ij}(u_i - u_j)^2 \quad \text{subject to} \quad \left||A| - |\overline{A}|\right| \leq 1, \\
= \quad & \frac{1}{2}\mathbf{u}'(\mathbf{D} - \mathbf{Y})\mathbf{u} \qquad \text{subject to} \qquad \left||A| - |\overline{A}|\right| \leq 1, \\
= \quad & \frac{1}{2}\mathbf{u}'\mathbf{L}\mathbf{u}; \ \text{subject to} \ \left| \, |\mathbf{A}| - |\overline{\mathbf{A}}| \, \right| \leq \mathbf{1} \qquad (4) \\
= \quad & (\text{bipartitioning width})
\end{aligned}
$$

Problem Formulation

- Where $|A|$ represents the size of the subgraph A, that is the number of vertices in A; Cutsize($\mathbf{u}$) is independent of the diagonal of the adjacency matrix $\mathbf{Y}$; and the matrix $\mathbf{L = D}$ - $\mathbf{Y}$ is named graph Laplacian matrix.

- When the number of vertices, $|V| = c$, $| \, |A| - |\overline{A}| \, | \leq 1$ is an even integer then the constraint in (4) becomes $|A| = |\overline{A}|$.

Problem Formulation

- The matrix $\mathbf{L}$ is positive semi-definite, because it can be expressed as

$$\mathbf{L} = \mathbf{S}'\mathbf{S} \tag{5}$$

where $\mathbf{S} = (s_{ij})$ is the signed edge-vertex incidence matrix of size $n \times c$. The entry $s_{ij} = 1$ if v_j is a starting point of the edge e_i, and $s_{ij} = -1$ if v_j is an ending point of the edge e_i, and $s_{ij} = 0$ otherwise.
- For instance, for the FH dataset and any one of 19 signed edges $\{v_1, v_3\} = \{ToF, PS\}$ can be represented as a $0/-1/1$ signed incidence indicator vector of size 13

$$(1 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0).$$

Problem Formulation

Table 3: The L Matrix.

	ToF	*VSD*	PS	*TGV*	PDA	AS	ASD	Tru	TA	CoA	Dex	Ptr	A-V
ToF	52	-13	-19	-10	-4	-1	-1	0	-1	0	-1	-2	0
VSD	-13	37	-3	-5	-3	-3	-6	-1	0	0	-2	-1	0
PS	-19	-3	30	-2	0	-1	-1	-3	-1	0	0	0	0
TGV	-10	-5	-2	26	-4	-1	-2	-1	0	-1	0	0	0
PDA	-4	-3	0	-4	17	-2	0	-1	-2	0	0	0	-1
AS	-1	-3	-1	-1	-2	16	-2	0	-1	-3	-2	0	0
ASD	-1	-6	-1	-2	0	-2	15	0	-1	-1	0	0	-1
Tru	0	-1	-3	-1	-1	0	0	7	0	0	0	-1	0
TA	-1	0	-1	0	-2	-1	-1	0	6	0	0	0	0
CoA	0	0	0	-1	0	-3	-1	0	0	5	0	0	0
Dex	-1	-2	0	0	0	-2	0	0	0	0	5	0	0
Ptr	-2	-1	0	0	0	0	0	-1	0	0	0	4	0
A-V	0	0	0	0	-1	0	-1	0	0	0	0	0	2

Relaxation

- Given that equation (4) is nondeterministic polynomial (NP) complete, we solve its continuous relaxed version,

$$\min_{\mathbf{u}\in\mathbf{R}^c} \; \mathit{RelaxedCutsize}(\mathbf{u}) = \frac{1}{2}\mathbf{u}'\mathbf{L}\mathbf{u}; \text{ subject to } \mathbf{u}'\mathbf{M}\mathbf{u} = \tau \quad (6)$$

  where M is a positive definite matrix and $\tau = (11\ldots1)'\mathbf{M}(11\ldots1)$.

- The solution for the above equation is given by the generalized eigenvectors of the graph Laplacian matrix

$$\mathbf{L}\mathbf{u} = \mu\mathbf{M}\mathbf{u} \quad (7)$$

Relaxation

- There are other size loss functions depending on (3) similar to (4), such as, Ratio Cut, Normalized Cut, Min Max Cut and Cheeger Cut, see for instance Ding (2004) or von Luxburg (2007).
- When $\mathbf{M} = \mathbf{I}$, the vector $\mathbf{u}_1$ called Fiedler vector, and the corresponding eigenvalue μ_1, Fiedler eigenvalue, because Fiedler (1973) first associated μ_1 with the connectivity of the graph and suggested partitioning vertices by separating them according to the sign of entries in the corresponding eigenvector $\mathbf{u}_1$.
- When $\mathbf{M} = \mathbf{D}$ in (7), this corresponds to the spectral decomposition of the normalized graph Laplacian matrix, see for instance von Luxburg (2007), or to Lebart's (1969, 2000) Contiguity Analysis.

Correspondence Analysis

- Let $\mathbf{X}$ be $\mathbf{Z}$ or $\mathbf{W}$ of dimension $r \times c$, and $\mathbf{P} = \mathbf{X}/t_{\mathbf{X}}$ its correspondence matrix, where $t_{\mathbf{X}} = \sum_{j=1}^{c} \sum_{i=1}^{r} \mathbf{X}_{ij}$, the grand total of $\mathbf{X}$.
- Because of the principle of distributional equivalence, CA of $\mathbf{Z}$ is identical to CA of $\mathbf{W}$. So, we consider only CA of $\mathbf{Z}$, and designate its correspondence matrix by $\mathbf{P}_Z = \mathbf{Z}/t_Z = \mathbf{Z}/(2n) = \mathbf{Z}/y_{tot}$.
- The metric on the column space of $\mathbf{Z}$ is $\mathbf{D}_r = \mathbf{Diag}(p_{i.} = 2/y_{tot})$, and the metric on the row space of $\mathbf{Z}$ is $\mathbf{D}_c = \mathbf{Diag}(p_{.j} = degree(v_j)/y_{tot})$.

Correspondence Analysis

- Let $(\lambda_\alpha^Z, \varphi_\alpha^Z)$ for $\alpha = 0, \ldots, k = rang(\mathbf{Z}) - 1$ be the sequence of dispersion measures associated to the standard column coordinate vectors of CA of $\mathbf{Z}$, with $1 = \lambda_0^Z \geqslant \lambda_1^Z \geqslant \lambda_2^Z \ldots$ Then $(\lambda_\alpha^Z, \varphi_\alpha^Z)$ satisfy the eigenequation

$$\lambda_\alpha^Z \varphi_\alpha^Z = \mathbf{D}_c^{-1} \mathbf{P}_Z' \mathbf{D}_r^{-1} \mathbf{P}_Z \varphi_\alpha^Z, \text{ with } (\varphi_\alpha^Z)' \mathbf{D}_c \varphi_\beta^Z = \delta_{\alpha\beta}$$

  where $\delta_{\alpha\beta} = 1$ if $\alpha = \beta$ and $\delta_{\alpha\beta} = 0$ otherwise.
- Similarly consider CA of $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$ and its correspondence matrix $\mathbf{P}_B = \mathbf{B}/t_{\mathbf{B}}$.

Correspondence Analysis

- Let $(\lambda_\alpha^B, \varphi_\alpha^B)$ for $\alpha = 0, \ldots, k = rang(\mathbf{B}) - 1$ be the sequence of dispersion measures associated to the standard column coordinate vectors, with $1 = \lambda_0^B \geqslant \lambda_1^B \geqslant \lambda_2^B \ldots$ Then $(\lambda_\alpha^B, \varphi_\alpha^B)$ satisfy the eigenequation,

$$\lambda_\alpha^B \varphi_\alpha^B = \mathbf{D}_c^{-1} \mathbf{P}'_B \mathbf{D}_c^{-1} \mathbf{P}_B \varphi_\alpha^B, \text{ with } (\varphi_\alpha^B)' \mathbf{D}_c \varphi_\beta^B = \delta_{\alpha\beta}$$

- The total inertia of $\mathbf{Z}$ is defined to be

$$\begin{aligned} Total\ Inertia(\mathbf{Z}) &= \sum_{\alpha=1}^{c-1} \lambda_\alpha^Z \\ &= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - p_{i.} p_{.j})^2}{p_{i.} p_{.j}} \end{aligned}$$

PCA Matrix Norms

- In a series of papers, Choulakian (2003, 2005, 2006a, 2006b) developed PCA based on matrix norms, thus generalizing the classical PCA, or equivalently generalizing the well known singular value decomposition (SVD).

- The first surprising and unexpected result was that the centroid method of component analysis, which preceded Hotelling's PCA, is based on a particular matrix norm, named centroid matrix norm, see Choulakian (2003, 2005, 2006b).

Classical PCA

- Let $\mathbf{T}$ be a centered or standardized data set of dimension $I \times J$, where I observations are described by the J variables, that is, $\mathbf{T}'\mathbf{T}/I$ is the covariance or the correlation matrix.
- For a vector $\mathbf{u} \in \mathbf{R}^J$, we define its Euclidean or L_2-norm to be $||\mathbf{u}||_2 = (\mathbf{u}'\mathbf{u})^{\frac{1}{2}}$.
- Let $k = rank(\mathbf{T})$. The classical principal component analysis (PCA) consists of successive maximization of the variance or the square of the L_2-norm of the linear combination of the variables of the matrix $\mathbf{T}$ subject to a quadratic constraint; that is, it is based on the following optimization problem

$$max\, ||\mathbf{Tu}||_2 \quad \text{subject to} \quad ||\mathbf{u}||_2 = 1; \tag{8}$$

Classical PCA

- Equivalently, PCA can also be described as maximization of the square of the L_2-norm of the linear combination of the observations of the matrix $\mathbf{T}$

$$max \left|\left|\mathbf{T}'\mathbf{v}\right|\right|_2 \quad \text{subject to } ||\mathbf{v}||_2 = 1. \tag{9}$$

- Equation (8) is the dual of (9), and they can be reexpressed as matrix norms

$$\begin{aligned} \lambda_1 &= \max_{\mathbf{u}\in\mathbb{R}^J} \frac{||\mathbf{T}\mathbf{u}||_2}{||\mathbf{u}||_2}, \\ &= \max_{\mathbf{v}\in\mathbb{R}^I} \frac{||\mathbf{T}'\mathbf{v}||_2}{||\mathbf{v}||_2}, \\ &= \max_{\mathbf{u}\in\mathbb{R}^J,\mathbf{v}\in\mathbb{R}^I} \frac{\mathbf{v}'\mathbf{T}\mathbf{u}}{||\mathbf{u}||_2 ||\mathbf{v}||_2}. \end{aligned} \tag{10}$$

The solution to (10), λ_1, is the square root of the greatest eigenvalue of the matrix $\mathbf{T}'\mathbf{T}$ or $\mathbf{T}\mathbf{T}'$.

PCA based on matrix norms

Two important remarks are provided below,

- Historically, the centroid method of calculating principal components of a matrix $\mathbf{T}$ was based on the following optimization problem

$$\max \mathbf{u}'\mathbf{T}'\mathbf{T}\mathbf{u} = \max ||\mathbf{T}\mathbf{u}||_2^2 \text{ subject to } \mathbf{u} \in \{-1, +1\}^J,$$

  which represents the problem of maximizing a non negative quadratic form where the parameter values are discretized into -1 or 1.

- The taxicab decomposition of a matrix $\mathbf{T}$ is based on the following optimization problem

$$\max ||\mathbf{T}\mathbf{u}||_1 \text{ subject to } \mathbf{u} \in \{-1, +1\}^J, \tag{11}$$

  which is a well known and much discussed matrix norm related to Grothendieck problem, see for instance, Alon and Naor (2006).

Taxicab PCA

- TPCA consists of maximizing the L_1-norm of the linear combination of the variables of the matrix $\mathbf{T}$; more precisely, it is based on the following optimization problem

$$max \, ||\mathbf{Tu}||_1 \quad \text{subject to} \quad ||\mathbf{u}||_\infty = 1; \tag{12}$$

- Equivalently, TPCA can also be described as maximization of the L_1-norm of the linear combination of the rows of the matrix $\mathbf{T}$

$$max \, \left|\left|\mathbf{T}'\mathbf{v}\right|\right|_1 \quad \text{subject to} \quad ||\mathbf{v}||_\infty = 1. \tag{13}$$

Equation (12) is the dual of (13).

Taxicab PCA

- They can be reexpressed as matrix norms

$$\begin{aligned} \lambda_1 &= \max_{\mathbf{u}\in\mathbb{R}^J} \frac{||\mathbf{Tu}||_1}{||\mathbf{u}||_\infty}, \\ &= \max_{\mathbf{v}\in\mathbb{R}^I} \frac{||\mathbf{T}'\mathbf{v}||_1}{||\mathbf{v}||_\infty}, \\ &= \max_{\mathbf{u}\in\mathbb{R}^J,\mathbf{v}\in\mathbb{R}^I} \frac{\mathbf{v}'\mathbf{Tu}}{||\mathbf{u}||_\infty ||\mathbf{v}||_\infty}. \end{aligned} \tag{14}$$

- The solution to (14), λ_1, is given in (11). The first principal axes, $\mathbf{u}_1$ and $\mathbf{v}_1$, are defined as

$$\mathbf{u}_1 = \arg\max_{\mathbf{u}} ||\mathbf{Tu}||_1 \text{ such that } ||\mathbf{u}_1||_\infty = 1, \tag{15}$$

and

$$\mathbf{v}_1 = \arg\max_{\mathbf{v}} ||\mathbf{T}'\mathbf{v}||_1 \text{ such that } ||\mathbf{v}_1||_\infty = 1. \tag{16}$$

Taxicab Correspondence Analysis

- Often CA is identified as categorical PCA; that is, it is considered as an adaptation of PCA to contingency tables. Similarly, we consider TCA as an adaptation of TPCA to contingency tables.

- Here we introduce TCA of a contingency table $\mathbf{N} = (n_{ij})$ of two nominal variables with I rows and J columns. Let $\mathbf{P} = \mathbf{N}/n$ be the associated correspondence matrix with elements p_{ij} , where $n = \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij}$ is the sample size.

Taxicab Correspondence Analysis

- We define $p_{i.} = \sum_{j=1}^{J} p_{ij}$, $p_{.j} = \sum_{i=1}^{I} p_{ij}$, the vector $\mathbf{r} = (p_{i.}) \in \mathbb{R}^I$, the vector $\mathbf{c} = (p_{.j}) \in \mathbb{R}^J$, and $\mathbf{D}_r = Diag(\mathbf{r})$ a diagonal matrix having diagonal elements $p_{i.}$, and similarly $\mathbf{D}_c = Diag(\mathbf{c})$.
- The application of TPCA algorithm to $\mathbf{P}$, is named TCA of the contingency table $\mathbf{N}$. We put $\mathbf{P}_0 = \mathbf{P}$ and denote by $\mathbf{P}_\alpha$ be the residual correspondence matrix at the α-th iteration.
- We replace $\mathbf{T}$ by $\mathbf{P}$ and the numbering of the iterations α varies from 0 to k, where $k = rank(\mathbf{P}) - 1$.

Taxicab Correspondence Analysis

- For $\alpha = 0$, $\mathbf{P}_0 = \mathbf{P}$. Row and column profiles with their masses play an important role in both CA and TCA.
- Let $\mathbf{R}_0 = \mathbf{D}_r^{-1}\mathbf{P}_0 = (r_{ij}) = (p_{ij}/p_{i\cdot})$ designate the row profiles, that is for each i, $\sum_{j=1}^{J} r_{ij} = 1$.
- The cloud of row profiles with their masses is the set $\{(\mathbf{r}_{0i}, p_{i\cdot}) | \text{ for } i = 1, ..., I\}$, where $\mathbf{r}_{0i}$ is the ith row of $\mathbf{R}_0$.
- We will interpret the steps of TCA using row profiles, similar interpretation can also be done using column profiles where the cloud of column profiles with their masses is the set $\{(\mathbf{c}_{0j}, p_{\cdot j}) | \text{ for } j = 1, ..., J\}$, where $\mathbf{c}_{0j}$ is the jth row of $\mathbf{C}_0 = \mathbf{D}_c^{-1}\mathbf{P}_0'$.

Taxicab Correspondence Analysis

- For $\alpha = 0$, the optimization problem (12) becomes

$$\max ||\mathbf{P}_0\mathbf{u}||_1 = \max ||\mathbf{D}_r\mathbf{R}_0\mathbf{u}||_1 \quad \text{subject to} \quad ||\mathbf{u}||_\infty = 1 \qquad (17)$$
$$= \max \sum_{i=1}^{I} p_{i\cdot} \, |\mathbf{r}_{0i}\mathbf{u}| \quad \text{subject to} \quad ||\mathbf{u}||_\infty = 1.$$

- The objective function in (17) is the weighted L_1 dispersion of the projection of the row profiles $\mathbf{r}_{0i}$ on the axis $\mathbf{u}$. When $\alpha = 0$, the 0-th principal axes are, see (15) and (16)

$$\mathbf{u}_0 = \arg \max_{\mathbf{u} \in \{-1,+1\}^J} ||\mathbf{P}_0\mathbf{u}||_1 = \mathbf{1}_J \quad \text{and} \quad \mathbf{v}_0 = \arg \max_{\mathbf{v} \in \{-1,+1\}^I} \left|\left|\mathbf{P}_0{}'\mathbf{v}\right|\right|_1 = \mathbf{1}_I, \qquad (18)$$

where $\mathbf{1}_J$ is the J component vector with coordinates of 1's.

Taxicab Correspondence Analysis

- The 0-th principal factor scores are

$$\mathbf{f}_0 = \mathbf{D}_r^{-1}\mathbf{P}_0\mathbf{u}_0 = \mathbf{R}_0\mathbf{u}_0 = \mathbf{1}_I \quad \text{and} \quad \mathbf{g}_0 = \mathbf{D}_c^{-1}\mathbf{P}_0'\mathbf{v}_0 = \mathbf{C}_0\mathbf{v}_0 = \mathbf{1}_J, \tag{19}$$

- The corresponding principal axes are

$$\mathbf{u}_0 = sgn(\mathbf{g}_0) = \mathbf{1}_J \quad \text{and} \quad \mathbf{v}_0 = sgn(\mathbf{f}_0) = \mathbf{1}_I. \tag{20}$$

- And, the 0-th taxicab dispersion measure can be represented in many different ways as

$$\begin{aligned} \lambda_0 &= \left|\left|\mathbf{P}_0'\mathbf{v}_0\right|\right|_1 = ||\mathbf{p}_c||_1 = ||\mathbf{D}_c\mathbf{g}_0||_1 = \mathbf{u}_0'\mathbf{D}_c\mathbf{g}_0, \\ &= ||\mathbf{P}_0\mathbf{u}_0||_1 = ||\mathbf{p}_r||_1 = ||\mathbf{D}_r\mathbf{f}_0||_1 = \mathbf{v}_0'\mathbf{D}_r\mathbf{f}_0, \\ &= 1. \end{aligned} \tag{21}$$

Taxicab Correspondence Analysis

- When we repeat above procedure on the residual dataset, the first residual correspondence matrix becomes

$$\begin{aligned} \mathbf{P}_1 &= \mathbf{P}_0 - \mathbf{P}_0\mathbf{u}_0\mathbf{v}_0'\mathbf{P}_0/\lambda_0 \\ &= \mathbf{P}_0 - \mathbf{D}_r\mathbf{f}_0\mathbf{g}_0'\mathbf{D}_c/\lambda_0. \\ &= \mathbf{P} - \mathbf{p}_r\mathbf{p}_c'. \end{aligned} \tag{22}$$

- Note that $\mathbf{p}_r\mathbf{p}_c'$ represents the correspondence matrix under the assumption that the row and column variables are independent. This solution is considered trivial both in CA and in TCA.

Taxicab Correspondence Analysis

- For $\alpha = 1$, we define the residual row and column profiles to be: $\mathbf{R}_1 = \mathbf{D}_r^{-1}\mathbf{P}_1$ and $\mathbf{C}_1 = \mathbf{D}_c^{-1}\mathbf{P}_1$. In general, the α-th taxicab dispersion measure can be represented in many different ways

$$\begin{aligned} \lambda_\alpha &= ||\mathbf{P}_\alpha \mathbf{u}_\alpha||_1 = ||\mathbf{D}_r \mathbf{f}_\alpha||_1 = \mathbf{v}'_\alpha \mathbf{D}_r \mathbf{f}_\alpha, \\ &= \left|\left|\mathbf{P}'_\alpha \mathbf{v}_\alpha\right|\right|_1 = ||\mathbf{D}_c \mathbf{g}_\alpha||_1 = \mathbf{u}'_\alpha \mathbf{D}_c \mathbf{g}_\alpha. \end{aligned} \tag{23}$$

- And the $(\alpha+1)$-th residual correspondence matrix is

$$\begin{aligned} \mathbf{P}_{\alpha+1} &= \mathbf{P}_\alpha - \mathbf{D}_r \mathbf{f}_\alpha \mathbf{g}'_\alpha \mathbf{D}_c / \lambda_\alpha, \\ &= \mathbf{P}_0 - \sum_{\beta=1}^{\alpha} \mathbf{D}_r \mathbf{f}_\beta \mathbf{g}'_\beta \mathbf{D}_c / \lambda_\beta. \end{aligned} \tag{24}$$

Taxicab Correspondence Analysis

- One gets the data reconstitution formula both in TCA and CA as,

$$p_{ij} = p_{i.}p_{.j}\left[1 + \sum_{\alpha=1}^{k} f_\alpha(i)g_\alpha(j)/\lambda_\alpha\right]. \qquad (25)$$

- Similar to the ordinary CA, the total variability is defined to be $\sum_{\alpha=1}^{k}\lambda_\alpha^2$, and the proportion of the explained variation by the α-th principal axis is $\lambda_\alpha^2/\sum_{\beta=1}^{k}\lambda_\beta^2$, and the relative cumulative explained variation is

$$CEV(\alpha) = \sum_{\gamma=1}^{\alpha}\lambda_\gamma^2/\sum_{\beta=1}^{k}\lambda_\beta^2 \quad \text{for} \quad \alpha = 1, ..., k. \qquad (26)$$

Taxicab Correspondence Analysis

- The visual maps are obtained by plotting the points $(f_\alpha(i), f_\beta(i))$ for $i = 1, ..., I$ or $(g_\alpha(j), g_\beta(j))$ for $j = 1, ..., J$, for $\alpha \neq \beta$.
- An important property of TCA and CA is that columns (or rows) with identical profiles (conditional probabilities) receive identical factor scores.
- One important advantage of TCA over CA is that it stays as close as possible to the original data: It directly acts on the correspondence matrix $\mathbf{P}$ without calculating a dissimilarity (or similarity) measure between the rows or columns.
- TCA does not admit a distance interpretation between profiles; there is no chi-square like distance in TCA. Fichet (2009) described it as a scoring method.

Taxicab Correspondence Analysis of Z

- Applying TCA to the (0,1) table **Z** or **W**, we have the following new result
  Theorem: Along the first principal axis, the projected edges in TCA of **Z** or **W** will be clustered and the number of cluster points is less than or equal to 3.
- This theorem shows that TCA automatically clusters the edges of **Z** or **W** into at most 3 clusters on the first principal axis. This is an important feature as TCA of **Z** provides a robust discrete relaxed solution to the graph partitioning problem. The above theorem is similar to Theorem 3 in Choulakian (2008b).

Taxicab Correspondence Analysis of B

- Applying TCA to the matrix $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$, which is a positive definite matrix, we see that TCA of $\mathbf{B}$ is equivalent to the neglected centroid method, and it is different from TCA of $\mathbf{Z}$.
- Furthermore, Choulakian (2006b, 2008b) showed that TCA of $\mathbf{B}$ is less robust than TCA of $\mathbf{Z}$.
- Moreover, it is noticed in the analysis of real dataset on graph partitioning, that the first dimension of TCA of $\mathbf{B}$ looks like a discretization of the factor scores of the first dimension of CA of $\mathbf{Z}$ or $\mathbf{B}$. Based on these facts, we will consider in the sequel only TCA of the $\mathbf{Z}$ table.

Analysis of FH dataset

- We applied CA and TCA to the **W** matrix described in the introduction. It is important to note that this dataset has 3 influential entries: the proportion of the first 3 weights, (13+19+10)/111, is larger than 1/3 of the volume of the graph.
- Table 2 displays the dispersion measures, λ_α^2 and λ_α^Z and the associated cumulative explained variation, $CEV(\alpha)$ in %, of TCA and CA for the first 6 principal dimensions. We clearly see that the first 2 dimensions of TCA of **Z** explain $35.49 - 24.35 = 11.14\%$ more than the first 2 dimensions of CA of **Z**.

Analysis of FH dataset

Table 2: Dispersion measures and cumulative proportions of explained dispersion of FH dataset.

	TCA of **Z**		CA of **Z**	
α	λ_α^2	$CEV(\alpha)$	λ_α^Z	$CEV(\alpha)$
1	0.4386	18.63	0.7249	13.18
2	0.3970	35.49	0.6145	24.35
3	0.3488	50.31	0.5631	34.59
4	0.2784	62.13	0.5535	44.65
5	0.2241	71.65	0.5307	54.30
6	0.1977	80.05	0.5097	63.57

Analysis of FH dataset

- Figures 1 and 2 displays the visual maps of the 13 defects produced by CA and TCA of **Z** table. It is well known that visual maps produced by CA are heavily influenced by light weight categories; this is the case in Figure 1, where the category AV is the rarest occurring defect with 2 counts.
- Such a thing does not seem to happen in TCA; in Figure 2 we clearly see four boxed clusters of categories; this should be interpreted with Figure 4 having 9 clusters of edges, four of them boxed and five circled. Note that each edge represents a pair of siblings.

Analysis of FH dataset

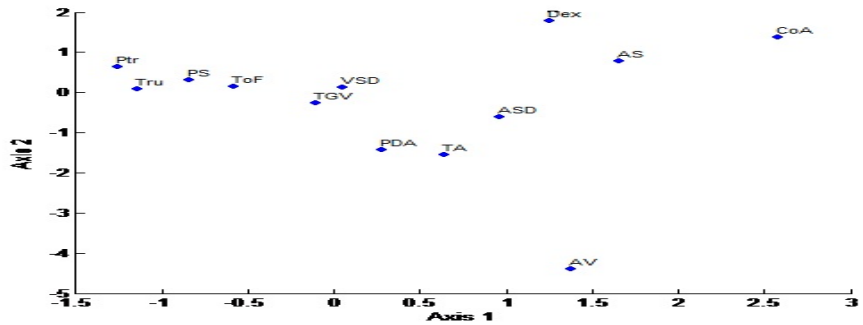

Figure: 1 FH dataset: CA of **Z** map of the 13 defects.

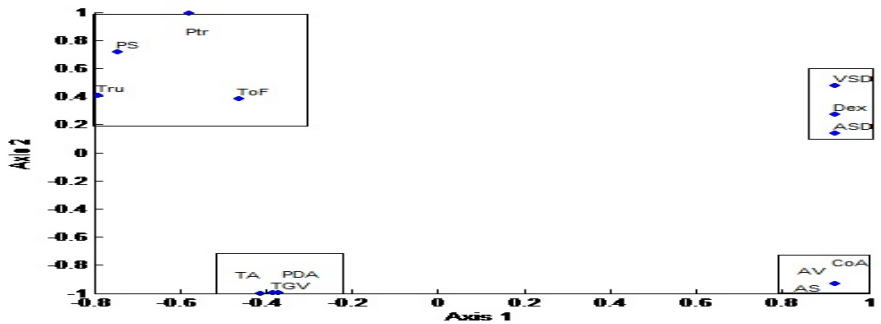

Figure: 2 FH dataset: TCA of **Z** map of the 13 defects.

Analysis of FH dataset

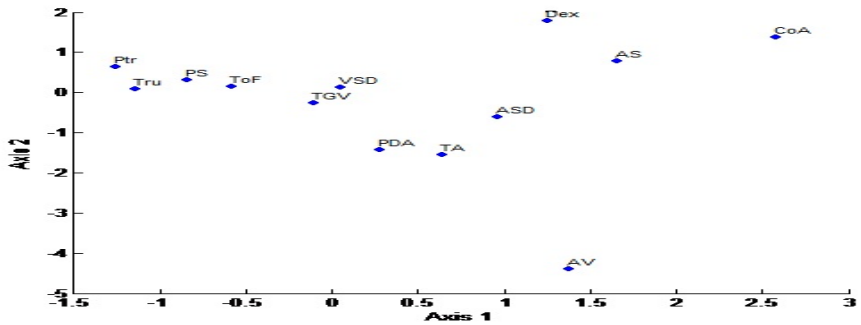

Figure: 1 FH dataset: CA of **Z** map of the 13 defects.

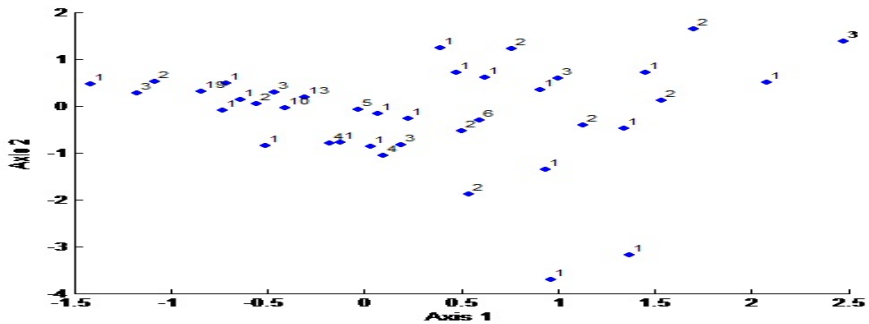

Figure: 3 FH dataset: CA of **Z** map of the edges with their frequencies.

Analysis of FH dataset

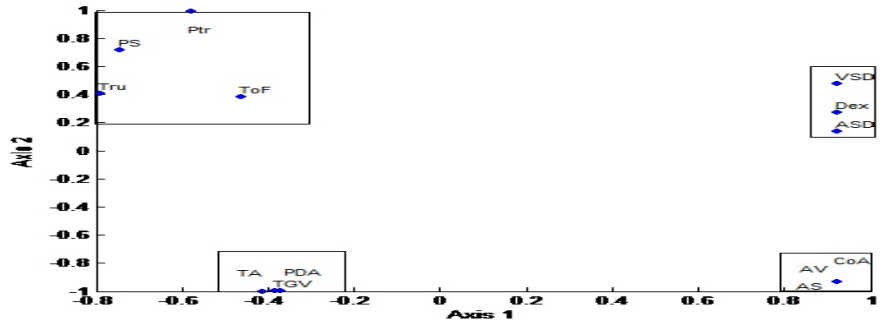

Figure: 2 FH dataset: TCA of **Z** map of the 13 defects.

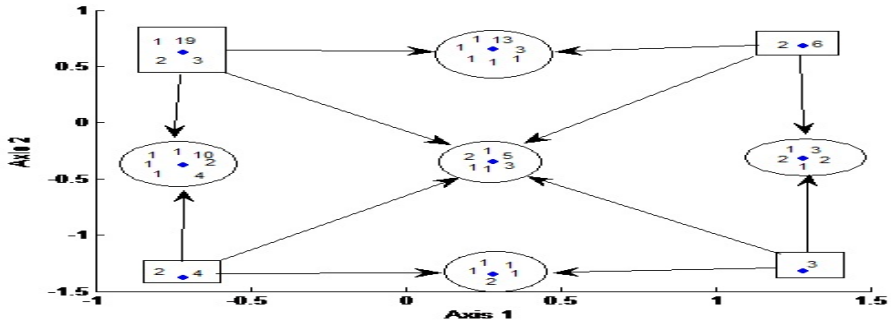

Figure: 4 FH dataset: TCA of **Z** map of the edges with their frequencies.

Analysis of FH dataset

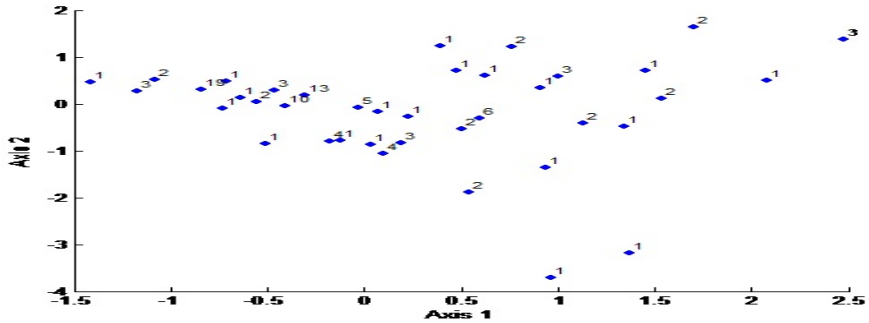

Figure: 3 FH dataset: CA of **Z** map of the edges with their frequencies.

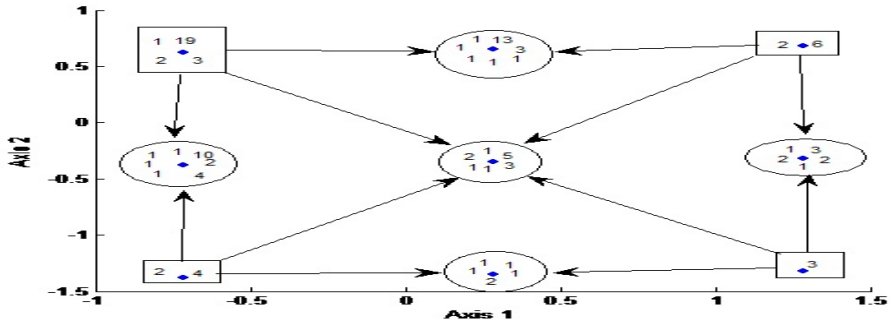

Figure: 4 FH dataset: TCA of **Z** map of the edges with their frequencies.

Analysis of FH dataset

- Figure 3 displays CA principal map of the 39 rows of the **W** matrix, where each row is identified by its frequency or count.
- The bottom two points identified by 1 represent the edges (AV, ASD) and (AV, PDA), refer to Table 1.
- Figure 4 displays TCA principal map of the 39 rows of the **W,** where we clearly see 9 clusters of points, four of them boxed and 5 circled.

Analysis of FH dataset

- Figures 2 and 4 complement each other and should be interpreted as: the four boxed clusters of variables in Figure 2 should be superimposed on the four boxed groups of edges in Figure 4. Each boxed cluster of edges in Figure 4 shows the number of intra edges within the associated boxed group of variables in Figure 2.
- For instance, the cluster of variables (VSD, Dex, ASD) found at the the upper right in Figure 2, will be associated with the 2 (VSD, Dex) and 6 (VSD, ASD) edges boxed at the upper right in Figure 4.
- Similarly, the cluster of variables (AV,CoA,AS) found at the the lower right in Figure 2, will be associated only with the 3 (CoA, AS) edges boxed at the lower right in Figure 4.

Analysis of FH dataset

- The circled edges are related by pointed arrows from the boxes; they represent inter-edges made up of paired siblings characterized from the associated groups of variables boxed in Figure 2.

- For instance, circled edges on the right of Figure 4 contain 1 (ASD,AV), 2 (ASD, AS), 1 (ASD, CoA), 3 (VSD,AS) and 2 (Dex,AS) paired siblings.

Analysis of FH dataset

- The first elements of the edges form the cluster of variables (ASD,VSD,Dex) which is found at the upper right corner of Figure 2, and the second elements of the edges form the cluster of variables (AV,AS,CoA) which is found at the lower right corner of Figure 2.

- The most heterogeneous groups of edges are circled in the middle of Figure 4, and which are related by four arrows coming from the four clusters of variables in Figure 2.

Conclusion

- First, we can interpret a symmetric contingency table with incomplete diagonals as the adjacency matrix of an unoriented graph with multiple edges, which can be coded as a 0/1 edge-vertex incidence data $\mathbf{Z}$, and to which CA or TCA can be applied.

- In this way one imputes values to the missing diagonal values, as suggested in Benzecri (1973, pages 240-241). Then CA will produce direct factors, and it will be equivalent to spectral decomposition of the graph Laplacian matrix which has a long history as old as CA.

Conclusion

- Second, TCA and CA of 0/1 matrices can produce different results, because the geometry underlying these two methods are different: TCA is based on the robust L_1 norm, while CA is based on the Euclidean norm.

- Third, the rows of 0/1 matrices in TCA will be clustered on the first principal axis and the number of clusters will be at most 3. We conjecture that the number of clusters of edges on the first principal plane will be at most 9; we do not have a formal proof of this conjecture.

Conclusion

- Fourth, visual maps produced by TCA and CA enrich each other: We have 2 different views of the data taken from 2 different angles; sometimes these views seem similar (second dataset) and other times dissimilar (first dataset).

- Finally, both methods provide relaxed solutions to the graph partitioning problem and complement each other: CA can detect exact block structures in a graph by the number of eigenvalues having values of 1; while TCA of $\mathbf{Z}$ can reveal the existence of balanced group structures in a graph.

References

BENZECRI, J.P. (1992).
Correspondence Analysis Handbook.
N.Y: Marcel Dekker.

CHOULAKIAN, V. (2008a).
"Taxicab Correspondence Analysis of Contingency Tables with One Heavyweight Column."
Psychometrika, 73: 309–319.

DE TIBEIRO, J. (1996).
"Sur les traits associés par paires : malformations cardiaques congénitales chez des enfants ayant mêmes parents."
Les Cahiers de l'Analyse des Données, 21, 45-52.

KOLACZYK, E.D. (2009).
Statistical Analysis of Network Data.
N.Y: Springer.

MOHAR, B. (1997).
"Some Applications of Laplace Eigenvalues of Graphs."
Graph Symmetry: Algebraic Methods and Applications, eds. Hahn, G. and Sabidussi, G., NATO ASI Ser. C 497, Kluwer, 225-275.