A Proposition for Fixing the Dimensionality of a Laplacian Low-rank Approximation of any Binary Data-matrix

Alain Lelu
Université de Franche-Comté/ELLIADD
LORIA
Campus Scientifique BP 239 - 54506
Vandoeuvre-lès-Nancy Cedex, France
alain.lelu@univ-fcomte.fr

Martine Cadot
Université de Lorraine
LORIA
Campus Scientifique BP 239 - 54506
Vandoeuvre-lès-Nancy Cedex, France
martine.cadot@loria.fr

***Abstract*— Laplacian low-rank approximations are much appreciated in the context of graph spectral methods and Correspondence Analysis. We address here the problem of determining the dimensionality K* of the relevant eigenspace of a general binary datatable by a statistically well-founded method. We propose 1) a general framework for graph adjacency matrices and any rectangular binary matrix, 2) a randomization test for fixing K*. We illustrate with both artificial and real data.**

Keywords-dimensionality reduction; intrinsic dimension; randomization test; low-rank approximation; graph Laplacian; bipartite graph; Correspondence Analysis; Cattell's scree; binary matrix.

I. INTRODUCTION AND STATE-OF-THE-ART

Spectral methods are used for optimally condensing and representing a set of objects in a space of lower dimensionality than the number of their descriptors. In this way, new relevant, informative and composite features are set apart from noisy, non-informative ones. This is especially useful when the descriptor space is sparse, which is typically the case for data of "pick-any" type, such as words in text segments, or links between Web pages, or social networks. An ever-growing number of applications rely on this type of condensed representation: Latent Semantic Analysis (LSA), supervised or semi-supervised learning, manifold learning, linear or non-linear PCA, vector symbolic architectures, spectral graph clustering and many others. A recurrent problem in any low-rank approximation process consists of determining the "right" rank of this approximation. Methods have been proposed to deal with this problem in the case of pre-defined data distributions, such as in [1]. But in the general case, nothing but empirical rules have been proposed, to the best of our knowledge: relying on the empirical evidence of a "gap" in the scree-plot of the eigenvalue sequence, whether visual or based on numerical indices such as first or second differences [2], or more basically on the value sqrt(N), etc. In LSA, empirical recommendations are provided [3], such as keeping the 200 to 400 first components. We address here the problem of determining the relevant dimensionality of the simplest, very common type of tabular data, i.e. the sparse binary tables. As such tables include adjacency matrices of unweighted graphs, and as graphs are known to be a powerful and extensively studied representation of many classes of data, we aim at incorporating in our framework a state-of-the-art representation space of graphs, i.e. one in the Laplacian family of eigenspaces. In the prospect of a maximum generality, a pleasant observation is that any binary datatable may be considered as a part of the adjacency matrix of a bipartite graph: we will focus, without lack of generality, on determining the best representation space, and its optimal number of dimensions, for unweighted and unoriented graphs, and thus for any binary matrix.

In Section II we will recall basic results about eigen-analysis of graphs and Correspondence Analysis. Section III will bridge the gap between graphs and general binary tables, and Section IV will present a randomization test for fixing the dimensionality of the relevant eigenspace. Applications to graph and data matrices are the topic of Section V, while we will close by presenting some related approaches, conclusions and future work.

II. EIGEN-SPACES FOR GRAPH MINING

To the best of our knowledge, the first application of eigen-analysis to graphs dates back to Benzécri [4], when Correspondence Analysis (C.A.) was applied to adjacency matrices. Let us recall that C.A. [5][6] relies on the eigen-analysis of a matrix **Q** issued from any two-way correspondence matrix **X** (in the case of an undirected and unweighted graph, **X** is binary and symmetric; **Q** is symmetric, too) with

$\mathbf{Q} = \mathbf{Dr}^{-1/2}\ \mathbf{X}\ \mathbf{Dc}^{-1/2}$,

where **Dr** and **Dc** are the diagonal matrices of the row and column totals. The eigen-decomposition of **Q** writes $\mathbf{Q} = \mathbf{U}\ \mathbf{\Lambda}\ \mathbf{V'}$ where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues (λ_1,..., $\lambda_L = 1$, L being the number of connected components; $1 > \lambda_{L+1} >... > \lambda_R > 0$, R being the rank of **X**). **U** and **V** are the eigenvector matrices for the rows and columns respectively, giving rise to several possible variants of C.A. factors, depending on the authors. Benzécri [4] has shown analytical solutions for simple graphs such as rings or meshes. Lebart [7] has generalized to contiguity analysis, and illustrated by showing that the (**F2, F3**) factor plane representation of the contiguity graph between French counties reconstitutes the allure of the France map.

 ISBN: 978-1-61208-254-7

An independent research track starting with [8] has defined two "normalized graph Laplacians", namely the symmetric Laplacian ($\mathbf{I} - \mathbf{Q}$), where $\mathbf{I}$ is the identity matrix, and $\lambda_1,..., \lambda_L = 0$, L being the number of connected components; $0< \lambda_{L+1}<...< \lambda_R$, R being the rank of $\mathbf{X}$. The "random walk" variant is $\mathbf{I} - \mathbf{D_r}^{-1} \mathbf{X}$.

Spectral graph clustering consists of grouping the similar nodes in a K-dimensional major eigen-subspace – for a review see [9] – and is an increasingly active research line. To our knowledge and up to now, the problem of determining the number K, when the distribution of degrees is non-standard, has not received more satisfactory answers than the scree-plot visual or second-difference heuristics [2], visually prominent in the case of small graphs, but difficult to put into practice in the case of large ones.

III. FROM GRAPHS TO BINARY MATRICES THROUGH BIPARTITE GRAPHS

A well-established result in data analysis states that the relevant, noise-filtered information lies in the dominant eigen-elements of a data matrix [8]. In the case of the $\mathbf{Q}$ matrix, Benzécri [4], Chung [8] and many others have shown that the value of its first eigenvalue, of multiplicity L (L being the number of connected components), is one (the same is true of the $\mathbf{D_r}^{-1} \mathbf{X}$ matrix).

In the case of a bipartite graph, whose adjacency matrix and symmetric Laplacian write respectively

$\begin{vmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M'} & \mathbf{0} \end{vmatrix}$ and $\begin{vmatrix} \mathbf{0} & \mathbf{Q} \\ \mathbf{Q'} & \mathbf{0} \end{vmatrix}$, a simplification follows

from the property of this type of matrices to have their eigenvalue set composed of the singular values of their rectangular non-empty submatrix, stacked with their opposites – in our Q case, in the range [-1; 1]. It follows that the basic correspondence analysis of any binary matrix, i.e. the SVD of its "symmetric" Laplacian matrix $\mathbf{Q}$, giving rise to the signed contributions of its rows and columns to the inertia accounted by each factor, constitutes a basic reference for comparing this matrix to random counterparts.

IV. A RANDOMIZATION TEST FOR FIXING THE DIMENSIONALITY OF THE RELEVANT EIGENSPACE

We have set up a randomization method [10] for generating random versions of a binary datatable with the same margins as those of this table, and set up the ensuing test for validating any statistics conducted on it. It is to be noted that the principles of generation of random matrices with same margins as a reference matrix seem to have been discovered independently several times, in various application domains: ecology, psychometrics, combinatorics, sociology. Cadot [11] legitimates a rigorous permutation algorithm based on rectangular "flip-flops", and shows that any Boolean matrix can be converted into any other one with the same margins in a finite number of cascading flip-flops, i.e. compositions of elementary rectangular flip-flops: at the crossings of rows i_1 and i_2, and columns j_1 and j_2, a rectangular flip-flop keeping the margins unchanged is possible if the (i_1, j_1) and (i_2, j_2) values are 1 whereas the (i_1, j_2) et (i_2, j_1) values are 0. To our knowledge it was the first time this principle was introduced in data mining.

As is the case for all other randomization tests [12], the general idea comes from the exact Fisher test [13], but it applies to the variables taken as a whole, and not pairwise. The flip-flops preserve the irreducible background structure of the datatable, but break up the meaningful links specific to a real-life data table. For example, most of texts×words datatables have a power-law distribution of the words, and a binomial-like one for the number of unique words in the texts. This background structure induces our "statistical expectation" of no links conditionally to the type of corpus. Getting rid of the background structure enables this method to process any type of binary data, both (1) taking into account the marginal distributions, (2) doing this without any need to specify any statistical model for these distributions.

When using this algorithm, one must fix the values of three parameters: the number of rectangular flip-flops for generating non-biased random matrices, the number of randomized matrices, the alpha risk. This test is akin to be applied to adjacency matrices of bipartite, unoriented, unweighted graphs, as the non-empty parts of such matrices are made up of two symmetric rectangular binary matrices, and this structure is akin to be reproduced when generating random versions as described above. For generating randomized versions of the adjacency matrix of an unoriented, unweighted graph, further constraints have to be imposed at the step of enabling or not a rectangular flip-flop: the square matrix must be kept symmetric and its diagonal empty. Note that the problem at stake is different from the one addressed by [19], i.e. generating a random matrix with prescribed margins, which does not necessarily supports an exact solution.

V. APPLICATIONS

We have detected the relevant dimension K* and used the corresponding reduced dataspace in the context of both graph and non-graph problems. In a proof-of-concept perspective, we will present here one artificial and one real dataset for each category – for a more detailed but less general presentation, see [14]. Throughout these examples, we will put forward successive visual representations we found useful.

A. Graphs

First, we have built the adjacency matrix of an unoriented and unweighted graph of 66 vertices, with four noisy cliques (missing intra-clique links: 17%; inter-clique links: 12%). Figure1 shows that the computed sequence of the 66 eigenvalues of its Laplacian, whether positive or negative, ranked by decreasing value of their module, fits into the "confidence funnel" of its 200 randomized counterparts, except the first one (value: one, by construction) and the next three, which clearly de-

 ISBN: 978-1-61208-254-7

lineate the relevant support space for representing the four clusters as vertices of a tetrahedron (Figure2).

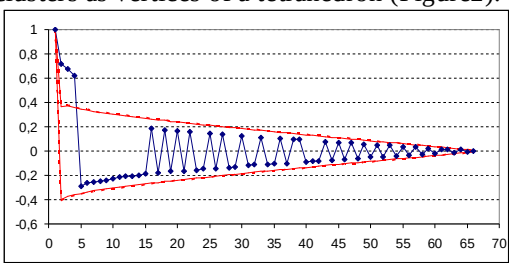

Figure 1. Graph with four noisy clusters: confidence funnel (red) of the eigenvalues (blue).

We have also processed the "Football league" data [15] which embed the "theoretical" social structure made of 12 regional "conferences", as well as the unsupervised structure emanating from the 115-node graph.

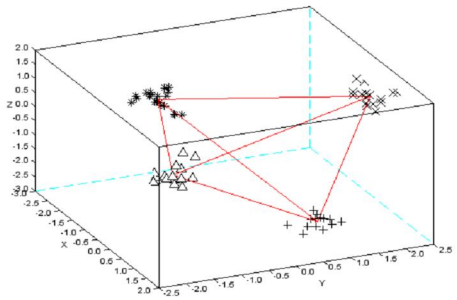

Figure 2. Graph with four noisy clusters: the 3-eigenvector relevant representation space of the 4 clusters.

According to our test with 200 randomized adjacency matrices, at the 99% confidence threshold, the ten "first" eigenvalues (N°2 to N°11, as there is a single connected component in the graph) of the original Laplacian matrix clearly dominate the confidence "funnel" of its 200 randomized counterparts.

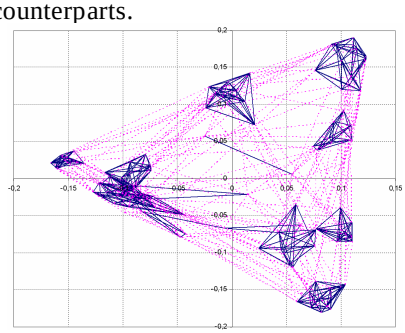

Figure 3. Football league: the U2 × U3 plane

An extra cluster analysis in this reduced space resulted in a quasi-perfect F-score measure for nine conferences, and a good or meager one for three of them, less geographically interrelated, summing up in a .956 global F-score. The U2 × U3 plane provides an overview of the structure (Figure 3), whereas the U4 to U11 dimensions offer more local points of view, and the subsequent ones show no recognizable structure.

B. Binary Rectangular Tables

We have designed a (1500; 836) binary matrix with a power-law distribution of the row sums and two fuzzy and overlapping column clusters, built by pasting twice the same (750; 836) Zipfian-distributed datatable, the second time with a random reordering of the columns. Our test results in two relevant eigenvalues, giving rise to a planar representation of the rows (Figure 4) showing off the orthogonality of two "logics", or "scales", and not a crisp opposition between two clusters.

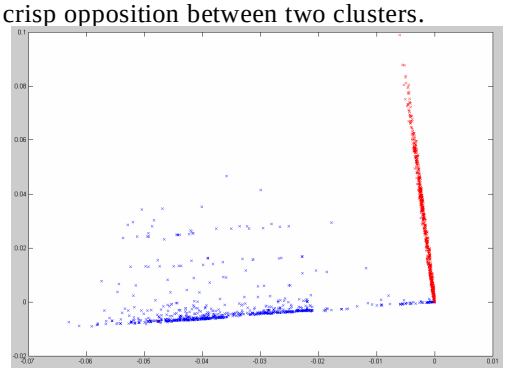

Figure 4. Data cloud with two overlapping "logics" projected onto the representation space of the two non-trivial and relevant eigenvectors U2 (vertical) and U1 (horizontal). In blue, the first 750 rows, in red the others.

Our experience is that this specific and rarely identified data structure is frequent in textual data; it takes here a concrete form when sorting the rows and columns of the datatable according to the dominant non-trivial eigenvector U2 (Figure 5).

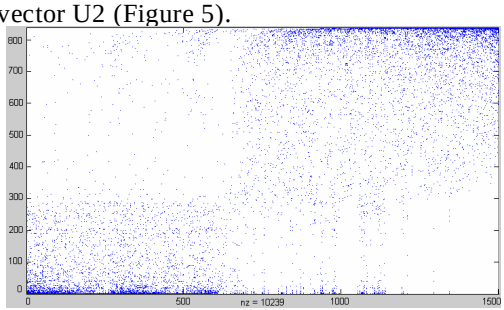

Figure 5. A binary datatable with two overlapping "logics" reordered by sorting the 1500 rows according to U2 and the 836 columns to V2.

We also processed a 753 per 11,567 Texts × Words matrix issued from a query to the Lexis-Nexis press database concerning three months of environmental controversies in the French press: it ensued that the relevant eigenspace was a 195-dimensional one, in which a cluster analysis of the words had put to the fore one clearly syntactical cluster from a hundred or so other ones devoted each to a particular press "story" noticeable in this period. In this case, the assessment criterion could be nothing but qualitative. Table 1 displays an example of such a news story.

 ISBN: 978-1-61208-254-7

TABLE I. EXAMPLE OF A CLUSTER IN THE PRESS DATABASE: THE NEWS STORY "THE EUROPEAN COMMISSIONER DACIAN CIOLOS NEGOTIATES AGRICULTURAL ISSUES IN WASHINGTON".

Rank	*Word*	*POS-tag*
1	antimicrobiens	U
2	spongiforme	U
3	Peterson	U
4	Ron	U
5	Mike	U
6	Dacian	U
7	CIOLOS	U
8	impression	SBC
9	09-févr	SBC
10	Lucas	U
11	US	SBC
12	durant	PREP
13	494	CAR
14	rappeler	PAR
15	préparation	SBC
16	répondre	PAR
17	conserver	VNCFF
18	subvention	SBC

VI. RELATED APPROACHES

While we have listed in Section I heuristic approaches for determining the relevant dimensionality of a data matrix, in [16] we presented a test in the same line as the one we develop here: we compared the singular values of a raw binary matrix to their counterparts in a set of randomized versions of this matrix. However this approach is subject to a major statistical concern: the singular-value scree does cross the upper bound of the singular-values of the randomized matrices, defining the desired relevant eigen-subspace, but it also crosses the lower bound, thus resulting in a difficult interpretation problem for the "significantly small" singular values. Moreover this approach offers no connection to Laplacian eigenspaces, nor Correspondence Analysis, as does the present one. Gionis et al. [17] deals, as we do, with the problem of finding out the number of relevant eigen-dimensions in a rectangular binary matrix, but presents a heuristic approach based on a unique randomized matrix, and no connection either to Laplacian eigenmaps nor Correspondence Analysis.

VII. CONCLUSION AND FUTURE WORK

We have presented a general framework for the dimensionality reduction of undirected and unweighted graphs, as well as of any rectangular binary table, a perspective covering both Laplacian eigenmaps and Correspondence Analysis of the said matrices. We have then shown that the number of dimensions of such an embedding space could be determined by a rigorous randomization test, contrasting with preceding heuristic approaches.

A major extension relates to scaling the procedure: whereas no efficiency issues arise for the data-class "n*1000 to m*10,000 vectors of n*1000 to m*10,000 dimensions", parallelization has to be set up beyond, both at the randomization level and the linear algebra computation one, which is well within the scope of the state-of-the-art. Another major extension, addressing both theoretical and practical difficult issues, is to generalize to any signed or unsigned integer matrix, if not any real-valued one.

ACKNOWLEDGMENTS

We thank the *CNRS/ISCC* for allowing us to use a part of its press data, and Azim Roussanaly for providing us with his efficient POS-Tagger [18].

REFERENCES

[1] Bouveyron C., Celeux G., and Girard S., "Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA", Statistics and Computing, vol. 17(4), 2007.

[2] Cattell R. B., "The scree test for the number of factors", Multivariate Behavioral Research, vol. 1(2), 1966, pp. 245–276

[3] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., and Harshman R. "Indexing by Latent Semantic Analysis". JASIS, vol. 41 (6), 1990, pp. 391–407 .

[4] Benzécri J.-P. L'analyse des données (3 tomes) Dunod, Paris, 1973

[5] Lebart L., Morineau A. and Warwick K., Multivariate Descriptive Statistical Analysis, John Wiley & sons, NY, 1984.

[6] Greenacre M., Correspondence Analysis In Practice, Chapman & Hall/crc Interdisciplinary Statistics Series, 2007.

[7] Lebart L., "Correspondence Analysis of Graph Structure", Comm. Meeting of the Psychometric Society, Bulletin Technique du CESIA, vol 2, 1984, pp. 5–19.

[8] Chung F.R.K., Spectral Graph Theory, (CBMS Regional Conference Series in Mathematics, No. 92), American Mathematical Society, 1997.

[9] Von Luxburg L., "A Tutorial on Spectral Clustering", Statistics and Computing, 2007, vol: 17(4).

[10] Cadot M., "A simulation technique for extracting robust association rules", CSDA'05, Chania, Greece, 2005.

[11] Cadot M., Extraire et valider les relations complexes en sciences humaines: statistiques, motifs et règles d'association. PhD thesis, Franche-Comté University, 2006.

[12] Manly B., Randomization, Bootstrap and Monte Carlo methods., Chapman and Hall/CRC, 1997.

[13] Fisher R., "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 1936, pp. 179–188.

[14] Lelu A. and Cadot M., "Espace intrinsèque d'un graphe et recherche de communautés", Revue I3, CEPADUES, Toulouse, 2011, vol. 11, pp. 1–25.

[15] Girvan M. and Newman M. E. J., "Community structure in social and biological networks", Proc. Natl. Acad. Sci. USA vol. 99, 2002, pp. 7821–7826.

[16] Lelu A., "Slimming down a high-dimensional binary datatable: relevant eigen-subspace and substantial content", COMPSTAT'10, Paris, 2010.

[17] Gionis, A., Mannila, H., Mielikäinen, T., and Tsaparas, P., "Assessing data mining results via swap randomization", ACM Trans. Knowl. Discov. Data, 2007.

[18] Roussanaly, A., Morph POS-Tagger: www.loria.fr/~azim; accessed on 12/24/2012.

[19] Lancichinetti A. and Fortunato S., "Benchmark for testing community detection algorithms on directed and weighted graphs with overlapping communities", Physical Review. Vol. E 80, 2009.

 ISBN: 978-1-61208-254-7