Summary of the work entitled *Difference of convex optimization for data visualization*

1 State of the art

In the Big Data era, Data Visualization is an area of interest to specialists from a wide variety of disciplines, [10, 11, 17, 18]. The information managed must be processed and, what is even more important, understood. Data Visualization techniques arise to respond to this requirement by developing specific frameworks to depict complex data structures as easy-to-interpret graphics, [24, 29].

Mathematical Optimization has contributed significantly to the development of this area during recent years, see [9, 20, 25] and the references therein. Nowadays, complex datasets pose new challenges in order to visualize the data in such a way that patterns are captured and useful information is extracted. Special attention is paid to represent the underlying dissimilarity relationships that data may have. Classical dimensionality reduction techniques, such as Principal Component Analysis, [26], or Multidimensional Scaling (MDS), [3], have been customized to deal with more complex data structures, [1, 12], and to make the interpretability of the results easier via, for instance, sparse models, [6, 5, 13].

Apart from adapting existing methods, specific problems may call also for new approaches. For instance, in addition to the dissimilarity measure, the data may have attached a statistical variable, to be related with the size of each object in the graphical representation of the dataset, [15]. This is the case for geographical data, to be visualized on a map in which countries are resized according to, for instance, population rates, but maintaining the neighboring relationships of countries. This type of representations, known as cartograms, [30], leads to plots in which countries are replaced by geometrical objects, frequently circles or rectangles, while the neighborhood relationships and the size of the objects are sought to be well represented. A key issue is how such problems are expressed as optimization programs, and which optimization tools are available to cope with them. For uses of optimization applied to cartograms construction and related visualization frameworks we refer the reader to [4, 7, 15, 16, 19, 21, 27, 28] and references therein.

2 Summary

In this paper we present a new mathematical programming framework to build a visualization map, in which a set of N individuals are depicted as convex objects in a bounded region $\Omega \subset \mathbb{R}^n$, usually $n \leq 3$. These objects must have a volume proportional to a given statistical value associated with the individuals, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$, and they should be placed accordingly to a dissimilarity measure attached to the individuals, $\boldsymbol{\delta} = (\delta_{ij})_{i,j=1,\ldots,N}$. In order to locate the objects in Ω, a reference object $\mathcal{B}$ is used, to be translated and expanded. However, since our final goal is to obtain a visualization map which allows the analysts to understand the data they are working with, a criterion which somehow controls the appearance of the plot needs to be also considered. We will deal with this paradigm by focusing on how the objects are spread out over Ω.

Leaving aside the statistical values $\boldsymbol{\omega}$, the purpose of representing dissimilarities between individuals reminds to MDS, which aims to represent the dissimilarity between individuals as empirical distances between points in an unbounded space of lower dimension. Although our visualization model may seem very close to MDS, it has the special feature of representing in the bounded region Ω not only dissimilarities as distances between objects, but also the statistical measure $\boldsymbol{\omega}$ through the volumes of the objects in Ω. Our visualization tool is able to rescale the dissimilarities between the individuals and the statistical values associated to them to fit in Ω. Observe that fitting the objects into Ω may yield representations in which the objects intersect if their sizes are not small enough, but, on the other hand, too small objects obstruct the visualization of the statistical measure. Ideally the objects should be spread out across the visualization map. This aim will be also taken into account when modeling the problem.

In this paper, the construction of a visualization map with the three characteristics mentioned above is written as a global biobjective optimization problem with convex constraints. We show that the objective function of the aggregated problem can be expressed as a difference of convex (DC) function, and thus DC optimization tools can be used to solve the optimization program, [22]. Moreover, a generalization of the previous problem is presented, in which a visualization map with multiple reference objects, $\mathcal{B}_1, \ldots, \mathcal{B}_S$, is built. This approach yields to a mixed integer nonlinear program for which an alternating algorithm is proposed. We show that, assuming that the choice about which object represents each individual has been made, the optimization problem is reduced to the single reference object case, and thus DC tools are appropriated to solve it. On the other hand, supposing that the visualization map is built, namely the locations of the objects representing the individuals are known, the problem of deciding which reference object best represents each individual turns into a nonconvex binary quadratic program.

3 Contribution to the literature

The main contribution of this work is the development of a general mathematical framework to visualize individuals related through a dissimilarity measure, which have also attached a statistical variable. Whereas the existing techniques in the literature are usually problem-specific, since they exploit the nature of the data, and specifically designed to depict two-dimensional plots, our proposal is able to handle data of any kind. Different criteria have been proposed and formally stated to obtain visual appealing maps, such that the concept of penetration depth or the using of different shapes to represent the objects. Including a bounded region Ω in which the visualization map has to be built allows the user to control the plot appearance, for instance not to have distortions when showing it on a computer screen.

From the algorithmic point of view, the powerful DCA has been applied to handle the model. The fact of being able to obtain a specific class of DC decomposition for the objective function makes the running times be dramatically reduced. Such decomposition allows us to obtain an explicit expression of the optimal solution of the convex problems to be solved in each iteration of DCA, and thus, we do not need to use any optimization routine to solve them.

In addition, the methodology proposed in this paper has applications in fields others than Data Visualization, such as for instance, Location Analysis, [14, 2, 8], or Distance Geometry, [23].

References

[1] H. Abdi, L. J. Williams, D. Valentin, and M. Bennani-Dosse. STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):124–167, 2012.

[2] R. Blanquero, E. Carrizosa, and P. Hansen. Locating objects in the plane using global optimization techniques. *Mathematics of Operations Research*, 34(4):837–858, 2009.

[3] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications.* Springer, 2005.

[4] K. Buchin, B. Speckmann, and S. Verdonschot. Evolution strategies for optimizing rectangular cartograms. In N. Xiao, M.-P. Kwan, M.F. Goodchild, and S. Shekhar, editors, *Geographic Information Science*, volume 7478 of *Lecture Notes in Computer Science*, pages 29–42. Springer, 2012.

[5] E. Carrizosa and V. Guerrero. Biobjective sparse principal component analysis. *Journal of Multivariate Analysis*, 132:151–159, 2014.

[6] E. Carrizosa and V. Guerrero. rs-Sparse principal component analysis: A mixed integer nonlinear programming approach with VNS. *Computers & Operations Research*, 52:349–354, 2014.

[7] E. Carrizosa, V. Guerrero, and D. Romero Morales. Piecewise rectangular visualization maps: A large neighborhood seach approach. Technical report, IMUS, Sevilla, Spain, 2015.

[8] E. Carrizosa, M. Muñoz-Márquez, and J. Puerto. Location and shape of a rectangular facility in $\mathbb{R}^n$. Convexity properties. *Mathematical Programming*, 83(1-3):277–290, 1998.

[9] E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers & Operations Research*, 40(1):150–165, 2013.

[10] C. P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[11] J. Choo and H. Park. Customizing computational methods for visual analytics with big data. *IEEE Computer Graphics and Applications*, 33(4):22–28, 2013.

[12] T. F. Cox and M. A. A. Cox. *Multidimensional scaling.* CRC Press, 2000.

[13] V. De Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004.

[14] J.M. Díaz-Báñez, J. A. Mesa, and A. Schöbel. Continuous location of dimensional structures. *European Journal of Operational Research*, 152(1):22–44, 2004.

[15] M. Dörk, S. Carpendale, and C. Williamson. Visualizing explicit and implicit relations of complex information spaces. *Information Visualization*, 11(1):5–21, 2012.

[16] D. Dorling. Area cartograms: their use and creation. In *Concepts and Techniques in Modern Geography series no. 59.* University of East Anglia: Environmental Publications, 1996.

[17] K. Fountoulakis and J. Gondzio. Performance of first- and second-order methods for big data optimization. Technical Report ERGO-15-005, 2015.

[18] K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, 156(1):189–219, 2016.

[19] E. Gomez-Nieto, F. San Roman, P. Pagliosa, W. Casaca, E. S. Helou, M. C. F. de Oliveira, and L. G. Nonato. Similarity preserving snippet-based visualization of web search results. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):457–470, 2014.

[20] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming*, 79(1-3):191–215, 1997.

[21] R. Heilmann, D. A. Keim, C. Panse, and M. Sips. Recmap: Rectangular map approximations. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 33–40. IEEE Computer Society, 2004.

[22] H.A. Le Thi and T. Pham Dinh. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1-4):23–46, 2005.

[23] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino. Euclidean distance geometry and applications. *SIAM Review*, 56(1):3–69, 2014.

[24] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[25] S. Olafsson, X. Li, and S. Wu. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429–1448, 2008.

[26] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[27] R. L. Rabello, G. R. Mauri, G. M. Ribeiro, and L. A. N. Lorena. A clustering search metaheuristic for the point-feature cartographic label placement problem. *European Journal of Operational Research*, 234(3):802–808, 2014.

[28] B. Speckmann, M. van Kreveld, and S. Florisson. A linear programming approach to rectangular cartograms. In *Proceedings of the 12th International Symposium on Spatial Data Handling*, pages 527–546. Springer, 2006.

[29] J. Thomas and P.C. Wong. Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

[30] W. Tobler. Thirty five years of computer cartograms. *Annals of the Association of American Geographers*, 94(1):58–73, 2004.