Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

A Study of the Optimality of PCA under Spectral Sparsification

Sergio Mercado [1]
Facultad Politécnica - UNA, Asunción - Paraguay
Marcos Villagra [2]
Facultad Politécnica - UNA, Asunción - Paraguay

1 Introduction

Principal component analisys (PCA) is a data analysis technique for mapping points in $\mathbb{R}^n$ to a two or three dimensional space. This dimensionality reduction preserves the natural grouping of points and information of data.

We say that data information is preserved if for all pair of distinct points $x, y \in \mathbb{R}^n$, the Euclidean distance $d(x, y)$ is similar to $d(\widetilde{x}, \widetilde{y})$, where $\widetilde{x}, \widetilde{y}$ are projections of x, y in two or three dimensional space. This is done optimally by an ortogonal projection of the points in $\mathbb{R}^n$ over the subspace generated by eigenvectors associated to the two or three greatests eigenvalues of the covariance matrix [1].

It is well known that computing eigenvalues in general is computationally expensive, and therefore, several authors use techniques of numerical approximation [3]. Furthermore, computations are more efficient whenever the matrices are sparse and memory costs can be reduced. It can be proved that adding zeros in a symmetric matrix M is equivalent to delete edges of a graph that represent M. This way, we can study this problem using graph theory.

2 Preliminaries

Spectral sparsification is a technique of algebraic graph theory that approximates an undirected weighted graph $G = (V, E, w)$ by a sparse subgraph (i.e. graph with fewer edges) $\widetilde{G} = (V, \widetilde{E}, \widetilde{w})$ under some approximation criteria. Many notions of approximation has been introduced, however, in this work we will use a notion of approxiation of Spielman and Teng [4] because it is a stronger notion than previous ones . Two graphs are similar if their respective quadratic Laplacian forms are. Formally, the Laplacian matrix of an

[1] sergio.mer.73@gmail.com
[2] mvillagra@pol.una.py

undirected weighted graph $G = (V, E, w)$, where $w(u, v)$ is the weight of the edge (u, v), is defined as

$$L_G(u,v) = \begin{cases} -w(u,v) & if \quad u \neq v \\ \\ \sum_z w(u,z) & if \quad u = v. \end{cases} \tag{1}$$

For all real vectors $x \in \mathbb{R}^n$, the *quadratic laplacian form* of a graph G in x is define as

$$x^T L_G x. \tag{2}$$

We say a graph $\widetilde{G}$ is a ϵ-spectral sparsifier of G if for all $x \in \mathbb{R}^n$ it holds that

$$(1-\epsilon)x^T L_{\widetilde{G}} x \leq x^T L_G x \leq (1+\epsilon)x^T L_{\widetilde{G}} x. \tag{3}$$

An important fact is that if $\widetilde{G}$ is a ϵ-spectral sparsifier of G, the eigenvalues of L_G and $L_{\widetilde{G}}$ are similar by an ϵ-factor. The best deterministic algorithm is of Zouzias [5], which constructs a spectral sparsifier of size $O(n/\epsilon^2)$ in $O(mn^2/\epsilon^2 + n^4/\epsilon^4)$ time. Lee and Sun [2] give the best probabilistic algorithm currently known.

3 Research Proposal

As an objective of this work we propose to study the optimality of PCA under a spectral sparsification process. To do this, we will generate a sparse matrix $\widetilde{S}$ that aproximes a covariance matrix S. By using the matrix $\widetilde{S}$ we will analyze the running time of PCA and the amount of information that is preserved.

Acknowledgement. S.M. is supported by FEEI- CONACyT-PROCIENCIA research grant PINV15-706 “COMIDENCO” and research grant POSG17-62. M.V. is supported by Conacyt research grant PINV15-208.

References

[1] I. Jolliffe, Principal component analysis. Springer Berlin Heidelberg, 2011.

[2] Y. T. Lee, and H. Sun. *Constructing linear-sized spectral sparsification in almost-linear time.* SIAM Journal on Computing, 47(6), 2315-2336., 2018.

[3] Y. Saad, Numerical methods for large eigenvalue problems: revised edition. Siam, 2011.

[4] D. A. Spielman, and S.H.Teng. *Spectral sparsification of graphs.* SIAM Journal on Computing, 40(4) : 981-1025, 2011.

[5] A. Zouzias. A matrix hyperbolic cosine algorithm and applications. In *Proceedings of the 39th International Colloquium on Automata, Languages, and Programming (ICALP)*, 2012.