Information Processing Letters 126 (2017) 35–38

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

NP-hardness and inapproximability of sparse PCA

Malik Magdon-Ismail

RPI CS Department, Troy, NY 12211, United States

ARTICLE INFO

Article history:
Received 27 February 2015
Received in revised form 25 May 2017
Accepted 26 May 2017
Available online 9 June 2017
Communicated by M. Chrobak

Keywords:
Combinatorial problems
Variable selection
NP-hard/Inapproximable
Sparsity constraint
Unsupervised learning

ABSTRACT

We give a reduction from CLIQUE to establish that sparse Principal Components Analysis (sparse PCA) is NP-hard. Using our reduction, we exclude a fully polynomial time approximation scheme (FPTAS) for sparse PCA (unless P=NP). Under stronger average case complexity assumptions, we also exclude polynomial constant-factor approximation algorithms.

1. Introduction

The earliest reference to principal components analysis (PCA) is in [15]. Since then, PCA has evolved into a classic tool for data analysis. A challenge for the interpretation of the principal components (or factors) is that they can be linear combinations of *all* the original variables. When the original variables have direct physical significance (e.g. genes in biological applications or assets in financial applications) it is desirable to have factors which have loadings on only a small number of the original variables. These interpretable factors are *sparse principal components (*SPCA*)*. There are many heuristics for obtaining sparse factors [3,19,20,7,6,13,17] as well as some approximation algorithms with provable guarantees [2, 4]. Our goal in this short paper is to establish the NP-hardness and inapproximability of SPCA using a reduction from CLIQUE.

The traditional formulation of sparse PCA is as cardinality constrained variance maximization:

Problem: SPCA (sparse PCA)
Input: Symmetric matrix $\mathrm{S} \in \mathbb{R}^{n\times n}$; sparsity $r \geq 0$; variance $M \geq 0$.
Question: Does there exist a unit vector $\mathbf{v} \in \mathbb{R}^n$ with at most r non-zero elements ($\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$ and $\|\mathbf{v}\|_0 \leq r$) for which $\mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v} \geq M$?

In the machine learning context, S is the covariance matrix for the data and, when there is no sparsity constraint, the solution $\mathbf{v}^*$ is the top right singular vector of S. A generalization of SPCA is the generalized eigenvalue problem for symmetric input matrices S and Q: maximize $\mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v}$ w.r.t. $\mathbf{v}$, subject to $\mathbf{v}^{\mathrm{T}}\mathrm{Q}\mathbf{v} = 1$ and $\|\mathbf{v}\|_0 \leq r$. This generalized eigenvalue problem is NP-hard [12] (via a reduction from sparse regression which is known to be NP-hard [14,8]). It is deeply embeded folklore that SPCA is NP-hard. The importance of sparse factors in dimensionality reduction has been recognized in some early work (the *varimax* criterion [10] has been used to rotate the factors to encourage sparsity, and this has been used in multi-dimensional scaling approaches to dimensionality reduction [16,11]).

E-mail address: magdon@cs.rpi.edu.

http://dx.doi.org/10.1016/j.ipl.2017.05.008
0020-0190/© 2017 Elsevier B.V. All rights reserved.

Summary of our results. We give a simple reduction from CLIQUE which shows that SPCA is NP-hard. This result also implies that it is NP-hard to determine the sign of the optimal objective value (if some algorithm $\mathcal{A}$ determined the sign of the optimal objective, one can solve SPCA by feeding $\mathrm{S} - M\mathrm{I}$ into $\mathcal{A}$). As we already mentioned, NP-hardness is folklore knowledge whose origin we are unable to determine. One of our contributions is to make this folklore a concrete fact together with a formal proof. The reduction from CLIQUE may be new. Our proof derives the input S to SPCA from the adjacency matrix of the input G to CLIQUE, so the problem remains NP-hard even if S is restricted to such inputs, whose entries are in $\{0, 1\}$. Further, instead of constructing S from the adjacency matrix of the input to CLIQUE, we can use the Laplacian, and our approach would still go through with some details changed. Thus, our results hold even when the input S is restricted to diagonally dominant matrices (a restriction of positive semi-definite). In typical machine learning applications, S is a covariance matrix which is positive semi-definite, and so SPCA remains NP-hard when restricted to that context.

Our main result is that there is no poly(n) $(1 - O(1/r^2))$-approximation algorithm for SPCA unless $P = NP$. This result also holds under the restrictions discussed above for NP-hardness. We should mention that subsequent related but independent work [4] studies the approximability of SPCA for positive semi-definite matrices, where they provide an $n^{-1/3}$-approximation and also show inapproximability to within $(1 - \epsilon)$ for some small ϵ.

Notation. $\mathrm{A}, \mathrm{B}, \ldots$ are matrices; $\mathbf{a}, \mathbf{b}, \ldots$ are vectors; and, $G, H, \ldots$ are graphs. The top eigenvalue of a matrix A is $\lambda_1(\mathrm{A})$; $\|\mathrm{A}\|_2$ is the spectral norm. For an undirected graph G, its adjacency matrix A is a (0,1)-matrix with $\mathrm{A}_{ij} = 1$ whenever edge (i, j) is in G. The spectral radius of a graph is the spectral norm of its adjacency matrix (also the top eigenvalue λ_1). $\mathbf{0}$ (resp. $\mathbf{1}$) are vectors or matrices of only zeros (resp. ones); for example, $\mathbf{1}_{m\times n}$ is a $m \times n$ matrix of ones.

2. Sparse PCA is NP-complete: reduction from CLIQUE

Problem: CLIQUE
Input: Undirected graph $G = (V, E)$; clique size K.
Question: Does there exist a K-clique in G?

The reduction is fairly straightforward. Given the inputs (G, K) for CLIQUE, we construct the inputs (S, r, M) for SPCA as follows. Let S be the adjacency matrix of G; let $r = K$; and, let $M = K - 1$. Clearly the reduction is polynomial. We now prove that there is a K-clique in G *if and only if* there is a K-sparse unit vector $\mathbf{v}$ for which $\mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v} \ge K - 1$. We need the following lemma on the spectral radius (top eigenvalue) of an adjacency matrix.

Lemma 1 *([5]). Let* A *be the adjacency matrix of a graph H of order ℓ. If H is an ℓ-clique, then $\|\mathrm{A}\|_2 = \lambda_1(\mathrm{A}) = \ell - 1$; if H is not an ℓ-clique, then $\|\mathrm{A}\|_2 = \lambda_1(\mathrm{A}) < \ell - 1$.*

We now prove the claim. Suppose Q is a K-clique in G and let S_Q be the $K \times K$ principal submatrix of S corresponding to the nodes in Q. Let $\mathbf{z}$ be a unit-norm top eigenvector of S_Q, and let $\mathbf{v}(\mathbf{z})$ be the vector with K non-zeros induced by $\mathbf{z}$: the non-zeros in $\mathbf{v}$ are at the indices corresponding to the nodes in Q and the values are the corresponding values in $\mathbf{z}$. Then,

$$\mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v} = \mathbf{z}^{\mathrm{T}}\mathrm{S}_Q\mathbf{z} = \lambda_1(\mathrm{S}_Q) = K - 1,$$

where the last equality follows from Lemma 1 because S_Q is the adjacency matrix of a K-clique. So, $\mathbf{v}(\mathbf{z})$ is a K-sparse unit vector for which $\mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v} \ge K - 1$. Now, suppose that there is a unit-norm K-sparse $\mathbf{v}$ for which $\mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v} \ge K - 1$. Let S_Q be the $K \times K$ principal submatrix of S corresponding to the non-zero entries of $\mathbf{v}$ and let $\mathbf{z}(\mathbf{v})$ be the K-dimensional vector consisting only of the non-zeros of $\mathbf{v}$. Let Q be the subgraph induced by the nodes corresponding to the non-zero indices of $\mathbf{v}$ (S_Q is the adjacency matrix of Q). Then, $\mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v} = \mathbf{z}^{\mathrm{T}}\mathrm{S}_Q\mathbf{z} \ge K - 1$, and so $\lambda_1(\mathrm{S}_Q) \ge K - 1$. By Lemma 1 if Q is not a K-clique then $\lambda_1(\mathrm{S}_Q) < K - 1$, so it follows that Q is a K-clique. Clearly SPCA is in NP and so it is NP-complete.

3. Inapproximability of SPCA

We now provide evidence that there is no efficient approximation algorithm for SPCA. First we rule out the possibility of a fully polynomial time approximation scheme (FPTAS). Given any instance (S, r) of SPCA, define $\mathsf{OPT}(\mathrm{S}, r) = \max_{\mathbf{v}} \mathbf{v}^{\mathrm{T}}\mathrm{S}\mathbf{v}$ over unit-norm r-sparse $\mathbf{v}$. A $(1 - \epsilon)$-approximation algorithm for SPCA produces a unit-norm r-sparse solution $\tilde{\mathbf{v}}$ for any given instance (S, r) satisfying $\tilde{\mathbf{v}}^{\mathrm{T}}\mathrm{S}\tilde{\mathbf{v}} \ge (1 - \epsilon)\mathsf{OPT}(\mathrm{S}, r)$. An FPTAS is algorithm to compute a $(1 - \epsilon)$-approximation for $\epsilon > 0$ and every instance of SPCA that is polynomial in n, r, ϵ^{-1}. The next theorem establishes that there is no polynomial $(1 - O(1/r^2))$-approximation algorithm and hence no FPTAS.

Theorem 2 *(No FPTAS). Unless P=NP, there is no polynomial time $(1 - \epsilon)$-approximation algorithm for* SPCA *with*

$$\epsilon < \epsilon^*(r) = \frac{r+1}{2(r-1)}\left(1 - \sqrt{1 - \frac{8}{(r+1)^2}}\right) = \frac{2}{r^2 - 1} + O(1/r^4).$$

The reason Theorem 2 implies no FPTAS is because if there were an FPTAS, then there would be a $(1 - O(1/r^2))$-approximation algorithm which runs in poly(n,r^2) time, which is poly(n) because $r \le n$. But Theorem 2 shows that there is no poly(n)-time $(1 - O(1/r^2))$-approximation algorithm.

Proof. The proof essentially amounts to strengthening Lemma 1 for the case that H is not an ℓ-clique. Specifically in Lemma 1, if adjacency matrix $\mathrm{A} \in \mathbb{R}^{\ell\times\ell}$ is not the adjacency matrix of an ℓ-clique, then we will show that

$$\begin{aligned}\lambda_1(\mathrm{A}) &\le \frac{\ell - 3}{2} + \frac{\ell + 1}{2}\left(1 - \frac{8}{(\ell+1)^2}\right)^{1/2} \\ &= (\ell - 1)(1 - \epsilon^*(\ell)). \qquad (*)\end{aligned}$$

Suppose that $(*)$ holds whenever H is not an ℓ-clique. For any SPCA instance (S, r), suppose the polynomial algorithm $\mathcal{A}$ gives a $(1-\epsilon)$-approximation with $\epsilon < \epsilon^*(r)$. We show how to use $\mathcal{A}$ to polynomialy decide CLIQUE. Given (G, K), the inputs to CLIQUE, use our reduction to construct $(\mathrm{S}, K, K-1)$, the inputs to SPCA. Now run algorithm $\mathcal{A}$ on (S, K) to obtain $\tilde{\mathbf{v}}$ and compute $x = \tilde{\mathbf{v}}\mathrm{S}\tilde{\mathbf{v}}$. If $x \geq (K-1)(1-\epsilon^*(K))$ then $\mathsf{OPT}(\mathrm{S}, K) = K-1$ and so there is a K-clique in G; if $x < (K-1)(1-\epsilon^*(K))$ then $\mathsf{OPT}(\mathrm{S}, K) < K-1$ (since we have a better than $(1-\epsilon^*(K))$-approximation) and so there is no K-clique in G.

To prove $(*)$, we first consider the adjacency matrix of a complete graph minus one edge,

$$\mathrm{A} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{1}_{2\times(\ell-2)} \\ \mathbf{1}_{(\ell-2)\times 2} & \mathbf{1}_{\ell-2}\mathbf{1}_{\ell-2}^{\mathrm{T}} - \mathrm{I}_{(\ell-2)\times(\ell-2)} \end{bmatrix}$$

By symmetry, the top eigenvector can be written $\begin{bmatrix} x\mathbf{1}_2 \\ y\mathbf{1}_{\ell-2} \end{bmatrix}$. The eigenvalue equation is

$$\begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{1}_{2\times(\ell-2)} \\ \mathbf{1}_{(\ell-2)\times 2} & \mathbf{1}_{\ell-2}\mathbf{1}_{\ell-2}^{\mathrm{T}} - \mathrm{I}_{(\ell-2)\times(\ell-2)} \end{bmatrix} \begin{bmatrix} x\mathbf{1}_2 \\ y\mathbf{1}_{\ell-2} \end{bmatrix} = \lambda \begin{bmatrix} x\mathbf{1}_2 \\ y\mathbf{1}_{\ell-2} \end{bmatrix},$$

and we obtain the equations:

$$(\ell-2)y = \lambda x;$$
$$2x + (\ell-3)y = \lambda y.$$

Solving for λ gives the quadratic $\lambda^2 - (\ell-3)\lambda - 2(\ell-2) = 0$, and the positive root is

$$\lambda = \frac{\ell-3}{2} + \frac{1}{2}\sqrt{(\ell+1)^2 - 8},$$

which is the expression in $(*)$. Since the spectral radius is strictly decreasing with edge-removal (using the Raleigh quotient and the Perron–Frobenius Theorem, see [18, page 9]), we have proved the upper bound in $(*)$. □

Under stronger (average-case) complexity assumptions we can also exclude polynomial constant factor approximations for SPCA. A natural optimization version of CLIQUE is the densest-K-subgraph (DκS): Given (G, K) find a subgraph Q on K nodes with the maximum number of edges. There is evidence that DκS does not admit efficient approximation algorithms [1].

Let G and G' be two graphs on n vertices. Suppose that one of the graphs has an ℓ-clique and for the other graph, every subgraph on ℓ vertices has at most $\delta\ell(\ell-1)/2$ edges for $0 < \delta < 1$. If one has a polynomial δ-approximation algorithm for DκS then one can determine which of G, G' has the ℓ-clique in polynomial time. We show that if one has an α-approximation algorithm for SPCA, then one can determine which of G, G' has the ℓ-clique in polynomial time for $\delta \leq \alpha^2$. This means that if there are no polynomial algorithms to distinguish between graphs with ℓ-cliques and graphs whose ℓ subsets are all below a density α^2, then there are no polynomial α-approximation algorithms for SPCA.

Suppose there is an α-approximation algorithm for SPCA. So, given any instance (S, r) of SPCA, in polynomial time one can construct a solution $\tilde{\mathbf{v}}$ for which $\tilde{\mathbf{v}}^{\mathrm{T}}\mathrm{S}\tilde{\mathbf{v}} \geq \alpha\mathsf{OPT}(\mathrm{S}, r)$. Let G, G' be the two graphs described above with $\delta = \alpha^2$. Note that

$$\delta = \alpha^2 < \alpha^2\frac{(\ell-1)}{\ell} + \frac{1}{\ell},$$

where the inequality is because $0 < \alpha < 1$. Now, let A be the adjacency matrix of G and run the α-approximation algorithm for SPCA with inputs (A, ℓ) to produce a solution $\tilde{\mathbf{v}}$. If $\tilde{\mathbf{v}}^{\mathrm{T}}\mathrm{A}\tilde{\mathbf{v}} \geq \alpha(\ell-1)$, declare that G contains the ℓ-clique; otherwise declare that G' contains the ℓ-clique. We prove that our algorithm correctly identifies the graph with the ℓ-clique.

If G does contain the ℓ-clique, then $\mathsf{OPT}(\mathrm{A}, \ell) = \ell - 1$ and the output $\tilde{\mathbf{v}}$ will satisfy $\tilde{\mathbf{v}}^{\mathrm{T}}\mathrm{A}\tilde{\mathbf{v}} \geq \alpha(\ell-1)$ (because it is an α-approximation) and so we will correctly identify G to have ℓ-clique. Now suppose that G does not contain the ℓ-clique. So, every ℓ-node subgraph in G has at most $e \leq \delta\ell(\ell-1)/2$ edges. We now use the bound on the spectral radius of a graph with e edges from [9]: $\|\mathrm{A}\|_2 \leq \sqrt{2e - n + 1}$, and since $e \leq \delta\ell(\ell-1)/2$, we have that

$$\begin{aligned} \|\mathrm{A}\|_2 &\leq \sqrt{\delta\ell(\ell-1) - \ell + 1} \\ &= \sqrt{\alpha^2\ell(\ell-1) - \ell + 1} \\ &< \sqrt{\left(\alpha^2\frac{(\ell-1)}{\ell} + \frac{1}{\ell}\right)\ell(\ell-1) - \ell + 1} \\ &= \alpha(\ell-1). \end{aligned}$$

Since $\|\mathrm{A}\|_2 < \alpha(\ell-1)$, we will correctly identify G' to have the ℓ-clique. The conclusion is summarized in the following theorem.

Theorem 3. *A polynomial α-approximation algorithm for* SPCA *gives a polynomial algorithm to distinguish between two graphs on n vertices, one of which contains an ℓ-clique and the other with every subset of ℓ nodes having at most $\alpha^2\ell(\ell-1)/2$ edges (for any (n, ℓ)).*

Under a variety of complexity assumptions it is known that one cannot efficiently distinguish between graphs with ℓ-cliques and graphs in which all subsets of size ℓ are sparse (for varying degrees of sparseness).

Theorem 4 *(No constant factor approximation for* DκS *[1]). Let $1 > \delta > 0$ be any constant approximation factor. Let G and G' be two graphs on ℓ^2 vertices. One of the graphs has an ℓ-clique and for the other graph, every subgraph on ℓ vertices has at most $\delta\ell(\ell-1)/2$ edges. Suppose there is no polynomial time algorithm for solving the hidden clique problem for a planted clique of size $n^{1/3}$. Then, there is no polynomial algorithm to determine which of G, G' has the ℓ-clique.*

Using Theorem 3 with Theorem 4,

Corollary 5 *(No constant factor approximation for* SPCA*). Suppose there is no polynomial time algorithm for solving* *the hidden clique problem for a planted clique of size* $n^{1/3}$*. Then, for any constant* $0 < \alpha < 1$*, there is no polynomial time* α*-approximation algorithm for* SPCA*.*

Acknowledgements

We thank Dimitris Papaillopoulos for suggesting reduction from CLIQUE. This research was partially supported by the Army Research Laboratory under Cooperative Agreement W911NF-09-2-0053 (the ARL-NSCTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] N. Alon, S. Arora, R. Manokaran, D. Moshkovitz, O. Weinstein, Inapproximability of densest κ-subgraph from average case hardness, Technical Report, www.tau.ac.il/~nogaa/PDFS/dks8.pdf.

[2] M. Asteris, D. Papailiopoulos, A. Dimakis, Non-negative sparse PCA with provable guarantees, in: Proc. ICML, 2014.

[3] J. Cadima, I. Jolliffe, Loadings and correlations in the interpretation of principal components, Appl. Stat. 22 (1995) 203–214.

[4] S.O. Chan, D. Papailliopoulos, A. Rubinstein, On the approximability of sparse pca, in: 29th Annual Conference on Learning Theory (COLT), vol. 49, 2016, pp. 623–646.

[5] L. Collatz, U. Sinogowitz, Spektren endlicher grafen, Abh. Math. Semin. Univ. Hamb. (1957) 63–77.

[6] A. d'Aspremont, F. Bach, L.E. Ghaoui, Optimal solutions for sparse principal component analysis, J. Mach. Learn. Res. 9 (June 2008) 1269–1294.

[7] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, G.R.G. Lanckriet, A direct formulation for sparse PCA using semidefinite programming, SIAM Rev. 49 (3) (2007) 434–448.

[8] D. Foster, H. Karloff, J. Thaler, Variable selection is hard, arXiv:1412.4832, 2014.

[9] Y. Hong, A bound on the spectral radius of graphs, Linear Algebra Appl. 108 (1988) 135–140.

[10] H. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (3) (1958) 187–200.

[11] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1) (1964) 1–27.

[12] B. Moghaddam, A. Gruber, Y. Weiss, S. Avidan, Sparse regression as a sparse eigenvalue problem, in: Proc. Information Theory and Applications Workshop (ITA), 2008.

[13] B. Moghaddam, Y. Weiss, S. Avidan, Generalized spectral bounds for sparse LDA, in: Proc. ICML, 2006.

[14] B. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. 24 (2) (1995) 227–234.

[15] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (1901) 559–572.

[16] J. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput. C-18 (5) (1969) 401–409.

[17] H. Shen, J.Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivar. Anal. 99 (July 2008) 1015–1034.

[18] D. Stevanovic, Spectral Radius of Graphs, Academic Press, 2014.

[19] N. Trendafilov, I.T. Jolliffe, M. Uddin, A modified principal component technique based on the lasso, J. Comput. Graph. Stat. 12 (2003) 531–547.

[20] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.