Optimal Average-Case Reductions to Sparse PCA: From Weak Assumptions to Strong Hardness

Matthew Brennan BRENNANM@MIT.EDU
Massachusetts Institute of Technology. Department of EECS.

Guy Bresler GUY@MIT.EDU
Massachusetts Institute of Technology. Department of EECS.

Editors: Alina Beygelzimer and Daniel Hsu

[1] Abstract

In the past decade, sparse principal component analysis has emerged as an archetypal problem for illustrating statistical-computational tradeoffs. This trend has largely been driven by a line of research aiming to characterize the average-case complexity of sparse PCA through reductions from the planted clique (PC) conjecture. All previous reductions either fail to show tight computational lower bounds matching existing algorithms or show lower bounds for formulations of sparse PCA other than its canonical generative model, the spiked covariance model. Also, these lower bounds all quickly degrade given weak forms of the PC conjecture. We give a reduction from PC that yields the first full characterization of the computational barrier in the spiked covariance model, providing tight lower bounds at all sparsities k. We also show the surprising result that even a mild improvement in the signal strength needed by the best known polynomial-time sparse PCA algorithms would imply that the hardness threshold for PC is $\text{polylog}(N)$, rather than on the order $N^{1/2}$ as is widely conjectured. This is the first instance of a suboptimal hardness assumption implying optimal lower bounds for another problem in unsupervised learning.

Keywords: statistical-computational gaps, average-case reductions, sparse PCA, planted clique conjecture, planted dense subgraph

1. Introduction

Principal component analysis (PCA), the task of projecting multivariate samples onto the leading eigenvectors of their empirical covariance matrix, is one of the most popular dimension reduction techniques in statistics. However in modern high-dimensional settings, PCA no longer provides a meaningful estimate of principal components Baik et al. (2005).

Sparse PCA was introduced in Johnstone and Lu (2004) to alleviate this inconsistency in the high-dimensional setting and has found applications in a diverse range of fields. In the past decade, sparse PCA has emerged as an archetypal problem for illustrating statistical-computational tradeoffs. This trend has largely been driven by a line of research (Berthet and Rigollet (2013b,a); Wang et al. (2016); Gao et al. (2017); Brennan et al. (2018)) working towards characterizing the average-case complexity of sparse PCA through reductions from the planted clique (PC) conjecture – which conjectures that there is no polynomial-time algorithm to detect a planted clique of size $K = o(N^{1/2})$ in $\mathcal{G}(N, \frac{1}{2})$. These reductions either fail to show tight computational lower bounds matching existing algorithms or show lower bounds for formulations of sparse PCA other than its

[1] Extended abstract. Full version available on arXiv with the same title.

canonical generative model, the spiked covariance model. Also, these lower bounds all quickly degrade with the exponent in the PC conjecture. When only given the PC conjecture up to $K = o(N^\alpha)$ where $\alpha < 1/2$, there is no sparsity k at which they remain tight. If $\alpha \leq 1/3$ these reductions fail to even show the existence of a statistical-computational gap at any sparsity k. Our main results are:

- We give a reduction from PC that yields the first full characterization of the computational barrier in the spiked covariance model, providing tight lower bounds at all sparsities k. This partially resolves a question raised in Brennan et al. (2018).

- We show the surprising result that weaker forms of the PC conjecture up to clique size $K = o(N^\alpha)$ for any given $\alpha \in (0, 1/2]$ imply tight computational lower bounds for sparse PCA at sparsities $k = o(n^{\alpha/3})$. Our reduction also shows that even a mild improvement in the signal strength needed by the best known polynomial-time sparse PCA algorithms would imply that the hardness threshold for PC is $\text{polylog}(N)$, rather than on the order $N^{1/2}$.

Our second result essentially shows that whether or not there are better efficient algorithms for PC is irrelevant to the statistical-computational gap for sparse PCA in the practically relevant highly sparse regime. This is the first instance of a suboptimal hardness assumption implying optimal lower bounds for another problem in unsupervised learning.

The reduction proving this result is more algorithmically involved than prior reductions to sparse PCA, making crucial use of a collection of average-case reduction primitives and introducing new techniques based on several decomposition and comparison properties of random matrices. Our lower bounds remain unchanged assuming hardness of planted dense subgraph instead of PC, which also has implications for the existence of algorithms slower than polynomial time. As a key step, our reduction maps an instance of PC to the empirical covariance matrix of sparse PCA samples, which proves to be a delicate task because of dependence among the entries of this matrix.

References

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, pages 1046–1066, 2013a.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013b.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *COLT*, pages 48–166, 2018.

Chao Gao, Zongming Ma, and Harrison H Zhou. Sparse CCA: adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *Unpublished manuscript*, 2004.

Tengyao Wang, Quentin Berthet, and Richard J Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.