# The Hong Kong University of Science and Technology

## Department of Mathematics

# <u>PhD THESIS EXAMINATION</u>

## Fast and Provable Algorithms for Sparse Principal Component Analysis

***By***


## Mr. Zhuozhi XIAN



***<u>ABSTRACT</u>***

Principal component analysis (PCA) is a well-known statistical method for feature extraction and dimension reduction used for data analysis. However, traditional PCA suffers from overfitting and loss of explainability in high-dimensional settings when variables exceed the sample size. Sparse PCA overcomes these issues by introducing sparsity into the principal components, yielding more interpretable results. In this thesis, we investigate the spiked covariance and spiked Wigner models in sparse PCA.

We first explore the spiked covariance model, which seeks to recover a sparse unit vector from noisy samples. Information theoretically, $\Omega(k \log p)$ observations suffice to recover a $k$-sparse $p$-dimensional vector $v$, yet existing polynomial-time methods require at least $\Omega(k^2)$ samples, revealing a gap. To bridge this, we introduce a novel thresholding-based algorithm that needs only $\Omega(k \log p)$ samples, provided the signal strength $\lambda = \Omega(\|v\|_\infty^{-1})$. We also propose a two-stage nonconvex algorithm that integrates our thresholding approach with truncated power iteration to achieve the minimax optimal rate of statistical error under the desired sample complexity. Numerical experiments validate the superior estimation accuracy and computational efficiency of our algorithms.

Secondly, we study the spiked Wigner model, which aims to recover an $s$-sparse $d$-dimensional unit vector $u$ from a $d \times d$ noisy matrix. The information theoretical lower bound for signal strength is $\beta = \Omega(\sqrt{s \log d})$, yet existing polynomial-time methods require at least $\widetilde{\Omega}(s)$, indicating a gap. To close it, we propose a new thresholding-based algorithm that demands only $\Omega(\sqrt{s \log d})$ signal strength, given $\|u\|_\infty = \Omega(1)$. We further design a two-stage nonconvex method combining our thresholding algorithm with truncated power iteration to achieve constant error in limited iterations. Empirical results demonstrate advanced performance in both estimation error and computational cost.


**Date : 7 Aug 2025, Thursday**

**Time : 10:00 am**

**Venue : Room 4472 (Lifts 25/26)**

***<u>Thesis Examination Committee</u>***:

| | | |
|---|---|---|
| **Chairman** | : | **Prof. Jun ZHANG, ECE/HKUST** |
| **Thesis Supervisor** | : | **Prof. Jianfeng CAI, MATH/HKUST** |
| **Member** | : | **Prof. Shing Yu LEUNG, MATH/HKUST** |
| **Member** | : | **Prof. Yang XIANG, MATH/HKUST** |
| **Member** | : | **Prof. Ruohan ZHAN, IEDA/HKUST** *(via online mode)* |
| **External Examiner** | : | **Prof. Song LI, School of Mathematical Science /Zhejiang University** |

***(Open to all faculty and students)***

**The student's thesis is now being displayed on the reception counter in the General Administration Office (Room 3461).**