# REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 19-05-2022 | Final Report | 1-May-2020 - 31-Oct-2021 |

4. TITLE AND SUBTITLE

Final Report: Trustworthy and Scalable Nonconvex Statistical Estimation for Sample-Starved Multi-Modal Data Models

5a. CONTRACT NUMBER
W911NF-20-1-0097

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Princeton University
PO Box 36
87 Prospect Avenue, Second Floor
Princeton, NJ 08544 -2020

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
75917-CS-YIP.10

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 15. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | YUXIN CHEN |
| UU | UU | UU | UU | | 19b. TELEPHONE NUMBER 215-898-8222 |

# RPPR Final Report
as of 23-Jun-2022

Agency Code: 21XD

Proposal Number: 75917CSYIP **Agreement Number: W911NF-20-1-0097**
**INVESTIGATOR(S):**


**Name:** YUXIN CHEN
**Email:** yuxin.chen@princeton.edu
**Phone Number:** 2158988222
**Principal:** Y

Organization: **Princeton University**
Address: PO Box 36, Princeton, NJ 085442020
Country: USA

DUNS Number: 002484665 EIN: 210634501
**Report Date:** 31-Oct-2021 Date Received: 19-May-2022
**Final Report** for Period Beginning 01-May-2020 and Ending 31-Oct-2021
**Title:** Trustworthy and Scalable Nonconvex Statistical Estimation for Sample-Starved Multi-Modal Data Models
**Begin Performance Period:** 01-May-2020 **End Performance Period:** 31-Oct-2021
**Report Term:** 0-Other

Submitted By: YUXIN CHEN Email: yuxin.chen@princeton.edu
Phone: (215) 898-8222

**Distribution Statement:** 1-Approved for public release; distribution is unlimited.

**STEM Degrees:** **STEM Participants:**

**Major Goals:** The overarching goal of this research program is to investigate reliable, model-agnostic, and provably accurate information processing and inference procedures in the challenging "data-hungry" regime (i.e. when the number of samples at hand is not necessarily much larger than the underlying degrees of freedom of the data models). Particular emphasis is placed on multi-modal and heterogeneous data models, including the cases where (1) the sensing units are heterogeneous so that the samples acquired might have drastically different characteristics, or (2) the acquired samples are
driven simultaneously by multiple important sources and we are asked to disentangle these sources. New insights and novel techniques from high-dimensional statistics, mathematical optimization, and statistical learning theory will be developed to meet the research objectives. Throughout the proposal, for concreteness, we frame our discussions in a few stylized problems such as mixed regression, blind deconvolution and de-mixing, spectral learning, tensor completion, etc. We believe that the techniques to be developed in this research program are broadly applicable to other foundational problems of critical values to the defense applications.

**Accomplishments:** During the past 1.5 years of the grant period, we have made progress towards multiple directions, as described below.

1. We have developed fast and guaranteed nonconvex algorithms showing how to effectively learn mixtures of low-rank models from random linear measurements. This is a challenging scenario that involves multi-modal and heterogeneous data.

2. We have developed statistically optimal inference procedures for estimating the principal subspace of a high-dimensional data stream, in the presence of heteroskedastic noise and missing data. The proposed procedures are fully data-driven and adaptive to heteroskedastic noise, without requiring prior knowledge about the noise levels and noise distributions.

3. We have developed efficient nonconvex optimization algorithms for low-rank tensor completion. We have also put forward statistically optimal uncertainty quantification algorithms that allow one to build entrywise confidence intervals for the unknown low-rank tensor.

4. We have established an intimate connection between convex relaxation and nonconvex optimization in robust principal component analysis and blind deconvolution, allowing one to demonstrate the statistical optimality of convex relaxation in the presence of random noise and adversarial outliers.

5. We have developed a suite of model-agnostic eigenvector inference procedures, which allow one to construct optimal confidence intervals for functions of eigenvectors. The procedures, which exploit how data asymmetry can help achieve bias reduction, work well without prior knowledge about the noise distributions and accommodate heterogeneous data models.

**Training Opportunities:** 9 PhD students and 2 postdoc have been partially supported by this grant, which provide valuable opportunities for research training for these students/postdoc.

**Results Dissemination:** PI Chen has presented more than 20 invited talks during the award period, disseminating the research results in multiple universities (e.g. Harvard, Stanford, UC Berkeley, Yale, CMU, UPenn, Caltech, RPI, USC, UIUC), workshops and conferences. In addition, PI Chen has coauthored a new monograph "Spectral Methods for Data Science: A Statistical Perspective", which has been published in the Foundations and Trends in Machine Learning series in 2021. PI Chen has also co-taught a tutorial in European Signal Processing Conference (EUSIPCO) 2020.

**Honors and Awards:** PI Chen has received the 2022 Alfred P. Sloan Research Fellowship under Mathematics, and the 2020 ICCM best paper award (gold medal). PI Chen has also received the 2021 Princeton SEAS junior faculty award.

**Protocol Activity Status:**

**Technology Transfer:** Nothing to Report

**PARTICIPANTS:**

**Participant Type:** Postdoctoral (scholar, fellow or other postdoctoral position)
**Participant:** Pengkun Yang
**Person Months Worked:** 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Jack Ji
**Person Months Worked:** 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Qinghua Liu
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Chengzhuo Ni
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)

**Participant:** Yuanhao Wang
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Yu Wu
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Yuling Yan
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Yuan Hui
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Yuren Zhong
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** PD/PI
**Participant:** Yuxin Chen
**Person Months Worked:** 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Postdoctoral (scholar, fellow or other postdoctoral position)
**Participant:** Gen Li
**Person Months Worked:** 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Igor Silin
**Person Months Worked:** 1.00 **Funding Support:**

# RPPR Final Report
as of 23-Jun-2022

Project Contribution:
National Academy Member: N

**Participant Type:** Graduate Student (research assistant)
**Participant:** Bingyan Wang
**Person Months Worked:** 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**ARTICLES:**

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
**Journal:** The Annals of Statistics
Publication Identifier Type: DOI Publication Identifier: 10.1214/20-AOS1986
Volume: 49 Issue: 2 First Page #:
Date Submitted: 5/29/21 12:00AM Date Published: 4/1/21 4:00AM
Publication Location:
**Article Title:** Subspace estimation from unbalanced and incomplete data matrices: $L_{2,\infty}$ statistical guarantees
**Authors:** Changxiao Cai, Gen Li, Yuejie Chi, H. Vincent Poor, Yuxin Chen
**Keywords:** spectral method, principal component analysis with missing data, tensor completion, covariance estimation, spectral clustering, leave-one-out analysis
**Abstract:** This paper is concerned with estimating the column space of an unknown low-rank matrix, given noisy and partial observations of its entries. There is no shortage of scenarios where the observations — while being too noisy to support faithful recovery of the entire matrix — still convey sufficient information to enable reliable estimation of the column space of interest. This is particularly evident and crucial for the highly unbalanced case where the column dimension d2 far exceeds the row dimension d1, which is the focal point of the current paper. We investigate an efficient spectral method, which operates upon the sample Gram matrix with diagonal deletion. While this algorithmic idea has been studied before, we establish new statistical guarantees for this method in terms of both ?2 and ?2,? estimation accuracy, which improve upon prior results if d2 is substantially larger than d1. To illustrate the effectiveness of our findings, we derive matching minimax lower bound
**Distribution Statement:** 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

# RPPR Final Report
as of 23-Jun-2022

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
**Journal:** IEEE Transactions on Information Theory
Publication Identifier Type: DOI Publication Identifier: 10.1109/TIT.2021.3050427
Volume: 67 Issue: 3 First Page #: 1928
Date Submitted: 5/29/21 12:00AM Date Published: 3/1/21 5:00AM
Publication Location:
**Article Title:** Nonconvex Matrix Factorization From Rank-One Measurements
**Authors:** Yuanxin Li, Cong Ma, Yuxin Chen, Yuejie Chi
**Keywords:** matrix factorization, rank-one measurements, gradient descent, nonconvex optimization
**Abstract:** We consider the problem of recovering low-rank matrices from random rank-one measurements, which spans numerous applications including covariance sketching, phase retrieval, quantum state tomography, and learning shallow polynomial neural networks, among others. Our approach is to directly estimate the low-rank factor by minimizing a nonconvex quadratic loss function via vanilla gradient descent, following a tailored spectral initialization. When the true rank is small, this algorithm is guaranteed to converge to the ground truth with near-optimal sample complexity and computational complexity. To the best of our knowledge, this is the first guarantee that achieves near-optimality in both metrics. In particular, the key enabler of near-optimal computational guarantees is an implicit regularization phenomenon: without explicit regularization, both spectral initialization and the gradient descent iterates automatically stay within a region incoherent with the measurements.
**Distribution Statement:** 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
**Journal:** The Annals of Statistics
Publication Identifier Type: DOI Publication Identifier: 10.1214/20-AOS1963
Volume: 49 Issue: 1 First Page #:
Date Submitted: 5/29/21 12:00AM Date Published: 2/1/21 5:00AM
Publication Location:
**Article Title:** Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices
**Authors:** Yuxin Chen, Chen Cheng, Jianqing Fan
**Keywords:** asymmetric matrices, eigenvalue perturbation, entrywise eigenvector perturbation, linear forms of eigenvectors, heteroscedasticity
**Abstract:** This paper is concerned with the interplay between statistical asymmetry and spectral methods. Suppose we are interested in estimating a rank-1 and symmetric matrix, yet only a randomly perturbed version M is observed. The noise matrix is composed of independent (but not necessarily homoscedastic) entries and is, therefore, not symmetric in general. This might arise if, for example, we have two independent samples for each entry of M? and arrange them in an asymmetric fashion. The aim is to estimate the leading eigenvalue and the leading eigenvector of the true matrix. We demonstrate that the leading eigenvalue of the data matrix M can be O (sqrt{n}) times more accurate (up to some log factor) than its (unadjusted) leading singular value of M in eigenvalue estimation. Moreover, the eigen-decomposition approach is fully adaptive to heteroscedasticity of noise, without the need of any prior knowledge about the noise distributions.
**Distribution Statement:** 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

# RPPR Final Report
as of 23-Jun-2022

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
**Journal:** IEEE Transactions on Information Theory
Publication Identifier Type: DOI Publication Identifier: 10.1109/TIT.2021.3065700
Volume: 67 Issue: 7 First Page #: 4613
Date Submitted: 10/28/21 12:00AM Date Published: 7/1/21 7:00AM
Publication Location:
**Article Title:** Learning Mixtures of Low-Rank Models
**Authors:** Yanxi Chen, Cong Ma, H. Vincent Poor, Yuxin Chena
**Keywords:** matrix sensing, latent variable models, heterogeneous data, mixed linear regression, non-convex optimization, meta-learning
**Abstract:** Spectral methods have emerged as a simple yet surprisingly effective approach for extracting information from massive, noisy and incomplete data. In a nutshell, spectral methods refer to a collection of algorithms built upon the eigenvalues (resp. singular values) and eigenvectors (resp. singular vectors) of some properly designed matrices constructed from data. A diverse array of applications have been found in machine learning, imaging science, financial and econometric modeling, and signal processing, including recommendation systems, community detection, ranking, structured matrix recovery, tensor data estimation, joint shape matching, blind deconvolution, financial investments, risk managements, treatment evaluations, causal inference, amongst others. Due to their simplicity and effectiveness, spectral methods are not only used as a stand-alone estimator, but also frequently employed to facilitate other more sophisticated algorithms to enhance performance.
**Distribution Statement:** 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 3-Accepted
**Journal:** Operations Research
Publication Identifier Type: DOI Publication Identifier: 10.1287/opre.2021.2106
Volume: Issue: First Page #:
Date Submitted: 6/12/21 12:00AM Date Published: 6/1/21 4:00AM
Publication Location:
**Article Title:** Nonconvex Low-Rank Tensor Completion from Noisy Data
**Authors:** Changxiao Cai, Gen Li, H. Vincent Poor, Yuxin Chen
**Keywords:** tensor completion, nonconvex optimization, gradient descent, spectral methods, entrywise statistical guarantees, minimaxity
**Abstract:** We study a noisy tensor completion problem of broad practical interest, namely, the reconstruction of a low-rank tensor from highly incomplete and randomly corrupted observations of its entries. Whereas a variety of prior work has been dedicated to this problem, prior algorithms either are computationally too expensive for large-scale applications or come with suboptimal statistical guarantees. Focusing on “incoherent” and well-conditioned tensors of a constant canonical polyadic rank, we propose a two-stage nonconvex algorithm—(vanilla) gradient descent following a rough initialization—that achieves the best of both worlds. Specifically, the proposed nonconvex algorithm faithfully completes the tensor and retrieves all individual tensor factors within nearly linear time, while at the same time enjoying near-optimal statistical guarantees (i.e., minimal sample complexity and optimal estimation accuracy).
**Distribution Statement:** 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

# RPPR Final Report
as of 23-Jun-2022

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
**Journal:** Foundations and Trends® in Machine Learning
Publication Identifier Type: DOI Publication Identifier: 10.1561/2200000079
Volume: 14 Issue: 5 First Page #: 566
Date Submitted: 10/28/21 12:00AM Date Published:
Publication Location:
**Article Title:** Spectral Methods for Data Science: A Statistical Perspective
**Authors:** Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma
**Keywords:** spectral methods; leave-one-out analysis
**Abstract:** Spectral methods have emerged as a simple yet surprisingly effective approach for extracting information from massive, noisy and incomplete data. In a nutshell, spectral methods refer to a collection of algorithms built upon the eigenvalues (resp. singular values) and eigenvectors (resp. singular vectors) of some properly designed matrices constructed from data. A diverse array of applications have been found in machine learning, imaging science, financial and econometric modeling, and signal processing, including recommendation systems, community detection, ranking, structured matrix recovery, tensor data estimation, joint shape matching, blind deconvolution, financial investments, risk managements, treatment evaluations, causal inference, amongst others. Due to their simplicity and effectiveness, spectral methods are not only used as a stand-alone estimator, but also frequently employed to facilitate other more sophisticated algorithms to enhance performance.
**Distribution Statement:** 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
**Journal:** IEEE Transactions on Information Theory
Publication Identifier Type: DOI Publication Identifier: 10.1109/TIT.2021.3111828
Volume: 67 Issue: 11 First Page #: 7380
Date Submitted: 10/28/21 12:00AM Date Published: 11/1/21 7:00AM
Publication Location:
**Article Title:** Tackling Small Eigen-Gaps: Fine-Grained Eigenvector Estimation and Inference Under Heteroscedastic Noise
**Authors:** Chen Cheng, Yuting Wei, Yuxin Chen
**Keywords:** Eigen-gap, linear form of eigenvectors, confidence interval, uncertainty quantification, heteroscedasticity
**Abstract:** Spectral methods have emerged as a simple yet surprisingly effective approach for extracting information from massive, noisy and incomplete data. In a nutshell, spectral methods refer to a collection of algorithms built upon the eigenvalues (resp. singular values) and eigenvectors (resp. singular vectors) of some properly designed matrices constructed from data. A diverse array of applications have been found in machine learning, imaging science, financial and econometric modeling, and signal processing, including recommendation systems, community detection, ranking, structured matrix recovery, tensor data estimation, joint shape matching, blind deconvolution, financial investments, risk managements, treatment evaluations, causal inference, amongst others. Due to their simplicity and effectiveness, spectral methods are not only used as a stand-alone estimator, but also frequently employed to facilitate other more sophisticated algorithms to enhance performance.
**Distribution Statement:** 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

# RPPR Final Report
as of 23-Jun-2022

**Publication Type:** Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
**Journal:** The Annals of Statistics
Publication Identifier Type: DOI Publication Identifier: 10.1214/21-AOS2066
Volume: 49 Issue: 5 First Page #:
Date Submitted: 5/19/22 12:00AM Date Published: 10/1/21 4:00AM
Publication Location:
**Article Title:** Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data
**Authors:** Yuxin Chen, Jianqing Fan, Cong Ma, Yuling Yan
**Keywords:** robust principal component analysis, nonconvex optimization, convex relaxation, leave-one-out analysis
**Abstract:** This paper delivers improved theoretical guarantees for the convex programming approach in lowrank matrix estimation, in the presence of (1) random noise, (2) gross sparse outliers, and (3) missing data. This problem, often dubbed as robust principal component analysis (robust PCA), finds applications in various domains. Despite the wide applicability of convex relaxation, the available statistical support (particularly the stability analysis vis-à-vis random noise) remains highly suboptimal, which we strengthen in this paper. When the unknown matrix is well-conditioned, incoherent, and of constant rank, we demonstrate that a principled convex program achieves near-optimal statistical accuracy. All of this happens even when nearly a constant fraction of observations are corrupted by outliers with arbitrary magnitudes. The key analysis idea lies in bridging the convex program in use and an auxiliary nonconvex optimization algorithm, a
**Distribution Statement:** 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

**CONFERENCE PAPERS:**

**Publication Type:** Conference Paper or Presentation **Publication Status:** 1-Published
**Conference Name:** International Conference on Machine Learning
Date Received: 29-May-2021 Conference Date: 12-Jul-2020 Date Published:
Conference Location: Virtual
**Paper Title:** Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality
**Authors:** Changxiao Cai, H. Vincent Poor, Yuxin Chen
Acknowledged Federal Support: **Y**

**Partners**

,

I certify that the information in the report is complete and accurate:
Signature: Yuxin Chen
Signature Date: 5/19/22 11:17PM

# ARO YIP Final Report (April 2020 - October 2021) Trustworthy and Scalable Nonconvex Statistical Estimation for Sample-Starved Multi-Modal Data Models

PI: Yuxin Chen (Princeton University)

During the past 1.5 years of the award period, we have made progress towards multiple directions, including learning nonconvex mixture models, uncertainty quantification for nonconvex statistical learning, and bridingconvex and nonconvex optimization. Several highlights are listed as follows.

- *Learning mixtures of low-rank models.* In this work, we consider the problem of learning mixtures of low-rank models, i.e., reconstructing a mixture of multiple low-rank matrices from unlabelled measurements of each. This problem enriches two widely studied settings — low-rank matrix sensing and mixed linear regression — by bringing latent variables (i.e., unknown labels) and structural priors (i.e., low-rank structures) into consideration. To cope with the non-convexity issues arising from unlabelled heterogeneous data and low-complexity structure, we develop a three-stage meta-algorithm that is guaranteed to recover the unknown matrices with near-optimal sample and computational complexities under Gaussian designs. In addition, the proposed algorithm is provably stable against random noise. We complement the theoretical studies with empirical evidence that confirms the efficacy of our algorithm.

- *Guaranteed nonconvex optimization for low-rank tensor completion.* This work is concerned with how to estimate a high-dimensional tensor with low CP-rank from highly incomplete and randomly corrupted measurements. Despite a flurry of recent activity in studying this problem, previous algorithms either are computationally too expensive, or suffer from statistical suboptimality. In order to address the inadequacy of past works, we propose an efficient two-stage nonconvex algorithm (i.e., first-order nonconvex method following spectral initialization) that provably achieves optimal statistical accuracy and computational effiency simultaneously. In particular, the proposed algorithm reconstructs the tensor and recovers all unknown tensor factors within nearly linear time, yielding an estimate with near-optimal statistical guarantees (i.e. minimal sample complexity and optimal estimation accuracy).

- *Uncertainty quantification for nonconvex tensor completion.* Moving from estimation to uncertainty quantification (an important step towards trustworthy decision making), we study the distribution and uncertainty of nonconvex optimization for noisy tensor completion. Focusing on the above-mentioned two-stage estimation algorithm, we characterize the distribution of our nonconvex estimator down to fine scales. This distributional theory in turn allows one to construct valid and short confidence intervals for both the unseen tensor entries and the unknown tensor factors. The proposed inferential procedure enjoys several important features: (1) it is fully adaptive to noise heteroscedasticity, and (2) it is data-driven and automatically adapts to unknown noise distributions. Furthermore, our findings unveil the statistical optimality of nonconvex tensor completion: it attains un-improvable Euclidean accuracy—including both the rates and the pre-constants—when estimating both the unknown tensor and the underlying tensor factors.

- *Bridging convex and nonconvex optimization in robust PCA and blind deconvolution.* In this work, we consider the convex programming approach in the problem of low-rank matrix estimation and that of blind deconvolution, in the presence of (1) random noise, (2) gross sparse outliers, and (3) missing data. These problems find applications in various application domains (e.g., channel estimation, recommendation systems). Despite the wide applicability of convex relaxation, the available statistical support (particularly the stability analysis vis-a-vis random noise) remains highly suboptimal, which

we strengthen in this paper. For a broad class of ground-truth signals, we demonstrate that a principled convex program achieves near-optimal statistical accuracy, in terms of both the Euclidean loss and the entrywise loss. All of this happens even when nearly a constant fraction of observations are corrupted by outliers with arbitrary magnitudes. All of this is enabled by bridging convex relaxation with the nonconvex Burer-Monteiro approach, a seemingly distinct algorithmic paradigm that is provably robust against noise and outliers. More specifically, we show that an approximate critical point of the nonconvex formulation serves as an extremely tight approximation of the convex solution, thus allowing us to transfer the desired statistical guarantees of the nonconvex approach to its convex counterpart.

- *Inference for heteroskedastic PCA with missing data.* This work studies how to construct confidence regions for principal component analysis (PCA) in high dimension, a problem that has been vastly under-explored. While computing measures of uncertainty for nonlinear/nonconvex estimators is in general difficult in high dimension, the challenge is further compounded by the prevalent presence of heteroskedastic noise and missing data. We propose a suite of solutions to perform valid inference on the principal subspace based on two estimators: a vanilla SVD-based approach, and a more refined iterative scheme called HeteroPCA. We develop non-asymptotic distributional guarantees for both estimators, and demonstrate how these can be invoked to compute both confidence regions for the principal subspace and entrywise confidence intervals for the spiked covariance matrix. Particularly worth highlighting is the inference procedure built on top of HeteroPCA, which is not only valid but also statistically efficient for broader scenarios (e.g., it covers a wider range of missing rates and signal-to-noise ratios). Our solutions are fully data-driven and adaptive to heteroskedastic random noise, without requiring prior knowledge about the noise levels and noise distributions.

- *Fine-grained eigenvector estimation and inference under heteroscedastic noise.* This work aims to address fundamental challenges arising in eigenvector estimation and inference for a low-rank matrix from *heterogeneous* noisy observations: (1) how to estimate an unknown eigenvector when the eigen-gap is particularly small; (2) how to perform estimation and inference on linear functionals of an eigenvector — a sort of "fine-grained" statistical reasoning that goes far beyond the usual $L_2$ analysis. Based on eigen-decomposition of the *asymmetric* data matrix (as well as certain data asymmetrization trick), we propose estimation and uncertainty quantification procedures for an unknown eigenvector in the presence of heterogeneous noise, which further allow us to reason about linear functionals of an unknown eigenvector. The proposed procedures and the accompanying theory enjoy several important features: (1) distribution-free (i.e. prior knowledge about the noise distributions is not needed); (2) adaptive to heteroscedastic noise; (3) minimax optimal under Gaussian noise.

**List of publications:**

1. "Spectral Methods for Data Science: A Statistical Perspective", Y. Chen, Y. Chi, J. Fan, C. Ma, *Foundations and Trends in Machine Learning*, vol. 14, no. 5, pp. 566-806, 2021.

2. "Inference for Heteroskedastic PCA with Missing Data", Y. Yan, Y. Chen, J. Fan, under revision, *Annals of Statistics*, 2022.

3. "Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise", C. Cheng, Y. Wei, Y. Chen, *IEEE Transactions on Information Theory*, vol. 67, no. 11, pp. 7380-7419, Nov. 2021.

4. "Learning Mixtures of Low-Rank Models", Y. Chen, C. Ma, H. V. Poor, Y. Chen, *IEEE Transactions on Information Theory*, vol. 67, no. 7, pp. 4613-4636, 2021.

5. "Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data", Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, *Annals of Statistics*, vol. 49, no. 5, pp. 2948-2971, Oct. 2021.

6. "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction", B. Li, S. Cen, Y. Chen, Y. Chi, *Journal of Machine Learning Research*, vol. 21, no. 180, pp. 1-51, 2020.

7. "Nonconvex Matrix Factorization From Rank-One Measurements", Y. Li, C. Ma, Y. Chen, Y. Chi, *IEEE Transactions on Information Theory*, vol. 67, no. 3, pp. 1928-1950, 2021.

8. "Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees", C. Cai, G. Li, Y. Chi, H. V. Poor, Y. Chen, *Annals of Statistics*, vol. 49, no. 2, pp. 944-967, 2021.

9. "Asymmetry helps: eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices", Y. Chen, C. Cheng, J. Fan, *Annals of Statistics*, vol. 49, no. 1, pp. 435-458, 2021.

10. "Convex and nonconvex optimization are both minimax-optimal for noisy blind deconvolution", Y. Chen, J. Fan, B. Wang, Y. Yan, accepted to *Journal of the American Statistical Association*, 2021.

11. "Nonconvex low-rank tensor completion from noisy data", C. Cai, G. Li, H. V. Poor, Y. Chen, accepted to *Operations Research*, 2020.

12. "Fast global convergence of natural policy gradient methods with entropy regularization", S. Cen, C. Cheng, Y. Chen, Y. Wei, Y. Chi, accepted to *Operations Research*, 2021.

13. "Softmax policy gradient methods can take exponential time to converge", G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, accepted to *Conference on Learning Theory (COLT)*, 2021.

14. "Uncertainty quantification for nonconvex tensor completion: confidence intervals, heteroscedasticity and optimality", C. Cai, H. V. Poor, Y. Chen, *International Conference on Machine Learning*, 2020.

15. "Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, *IEEE Transactions on Information Theory*, vol. 68, no. 1, pp. 448-473, Jan. 2022.

**Invited tutorial**

1. Y. Chen, Y. Chi, and C. Ma, "Nonconvex Optimization for High-Dimensional Signal Estimation: Spectral and Iterative Methods," European Signal Processing Conference (EUSIPCO) 2020.

**Award and honor**

1. Alfred P. Sloan Research Fellowship
2. Google Research Scholar Award
3. Coauthored a paper that wins the SLDS Student Paper Award, American Statistical Association (ASA)
4. Princeton SEAS Junior Faculty Award
5. ICCM Best Paper Award (Gold Medal)