REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
07-12-2017	Final Report	8-Sep-2014 - 7-Sep-2017

4. TITLE AND SUBTITLE
Final Report: Subspace Methods for Massive and Messy Data

5a. CONTRACT NUMBER
W911NF-14-1-0634

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
University of Michigan - Ann Arbor
3003 South State Street

Ann Arbor, MI 48109 -1274

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)
65692-MA.20

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			Laura Balzano
UU	UU	UU	UU		19b. TELEPHONE NUMBER 734-615-9451

RPPR Final Report
as of 08-Dec-2017

Agency Code:

Proposal Number: 65692MA **Agreement Number: W911NF-14-1-0634**
INVESTIGATOR(S):

Name: Laura Kathryn Balzano
Email: girasole@umich.edu
Phone Number: 7346159451
Principal: Y

Organization: **University of Michigan - Ann Arbor**
Address: 3003 South State Street, Ann Arbor, MI 481091274
Country: USA
DUNS Number: 073133571 EIN: 386006309
Report Date: 07-Dec-2017 Date Received: 07-Dec-2017
Final Report for Period Beginning 08-Sep-2014 and Ending 07-Sep-2017
Title: Subspace Methods for Massive and Messy Data
Begin Performance Period: 08-Sep-2014 **End Performance Period:** 07-Sep-2017
Report Term: 0-Other
Submitted By: Laura Balzano Email: girasole@umich.edu
Phone: (734) 615-9451
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 3 **STEM Participants:** 5

Major Goals: This proposal focused on subspace estimation in various modern big-data contexts, where data are massive, streaming, time-varying, and have missing, corrupted, and ill-conditioned data. In the final stages of the project, we also begun an exploration of nonlinear generalizations of subspaces, to provide models that can capture more interesting signal variation.

We formulated our general problem as solving for a factored representation for a data matrix X= UV', where columns of X are length n but X is rank r<n, given linear measurements from a measurement operator H. In this context we had five major goals:

1. Algorithms and theory for estimating U when X is ill-conditioned, i.e. the first and rth singular values have a large ratio, or when there is a low SNR, i.e. the r+1 to nth singular values are relatively large compared to the first r singular values. We also extended this in two new directions: a) clusters of data with different SNR levels, and b) data that are transformed by an unknown monotonic calibration function.

2. Theory for estimating U with the natural Grassmannian incremental gradient descent algorithm when H is a compressive measurement operator.

3. Algorithms and theory for the problem variant where we minimize the l1-norm cost instead of the Frobenius norm (robust PCA) or where we add a constraint that U should be sparse (sparse PCA).

4. Applications in computer vision and Unifying convergence theory across these non-convex problems.

5. Our added final goal was to develop algorithms and theory for estimating three nonlinear models with missing and compressive measurements: single index models, variety models, and union of subspace models.

Accomplishments: Goal 1) Subspace learning with ill conditioned, noisy data

Early in the grant period we developed PIMC, Polar Incremental Matrix Completion, which was the first online subspace learning algorithm that could handle ill-conditioned data.

We have developed first of their kind mathematical guarantees for convergence of the natural stochastic gradient algorithm for subspace estimation. We have global convergence guarantees with full (not compressive) measurements and noisy measurements (see comments on compressive measurements below). For noisy

measurements we derived a weighted step size scheme that gradually shrinks the step as the estimated subspace converges to the true subspace. The intuition behind this strategy is that the noisy part will gradually dominate the residual, and so we hope to include less and less information from the projection residual to our estimates over time. While this approach has been used in practice by others, we provide the first theoretical understanding of why the approach leads to good results on this nonconvex objective.

To the best of our knowledge, we are the first to study subspace estimation with heteroscedastic data, i.e. data with multiple noise levels, a very common messy data issue in real datasets. Our earlier results established precise formulas for the asymptotic behavior of PCA's estimates of the principal components (or left singular vectors). We also confirmed through simulation that our formulas are an exact prediction of PCA's performance when different data vectors have varying SNR levels. We continued this work in this final reporting period and have submitted to the Journal of Multivariate Analysis. We have significantly simplified the expression for our predictions of the principal components, and we have extended these to predict the behavior of the estimated singular values and right singular vectors. We have done an extensive comparison of our theory to classical high-dimensional PCA analysis approaches.

Another new direction for this work was to study whether low-rank matrices can be recovered even when all entries are transformed by an unknown monotonic function. This is common in problems such as recommender systems, where each individual's quantization function is unknown, and calibration of sensor systems, where calibration of each sensor drifts over time. We developed a first-of-its-kind algorithm for this purpose and demonstrated its performance empirically. This work focused on the low-dimensional subspace model and for Goal 5 we extended that to other simple structures, including sparse and group sparse, using the atomic set framework.

Goal 2) Compressive measurements

We have made progress on convergence results for both the case of compressive measurements as well as stronger results for missing data. We also had the objective of unifying convergence results across missing data contexts. This work has been in the review process since September 2016. We are in the process of a major revision, and will be resubmitting soon to the SIAM journal of optimization.

For undersampled data, we studied both missing data or compressively sampled data. Under mild conditions, we can prove that with high probability we have a unified convergence result as long as we have $m > O(d \log n)$ samples, where d is the subspace rank and n is the ambient dimension. These are impressive mathematical results, as the objective function of our formulation is nonconvex. We have also generalized the compressive measurement bounds considerably by considering sub-Gaussian measurement matrices in the case of subspace detection (not yet estimation). These are the first ever results for subspace learning that allow sub-Gaussian compressive operators, which are much more general than Gaussian compressive measurements.

We also considered compressive measurements for the subspace clustering problem. We showed that a Gaussian sketch is much better than directly subsampling, even with incoherent data, because it maximally separates the columns on the subsampled coordinates. This was previously unknown.

Goal 3) l1 norm subspace estimation

During the project we also worked on streaming sparse PCA algorithms. We developed both a batch algorithm as well as the first stochastic gradient algorithm for estimating sparse principal components; the latter works from streaming data. We showed the performance in simulation and on environmental ozone data. They key for these algorithms is in choosing the step size, and further understanding of our successful step size schemes is for future work.

Goal 4) Applications in CV and unification

We have applied our subspace tracking algorithms to the computer vision problems of structure from motion, face recognition, digit recognition, and motion segmentation in video.

Our work has demonstrated a general framework connecting the Grassmannian methods to the incremental Singular Value Decomposition. We've also unified convergence rates across complete data, noisy data, and undersampled data for our Grassmannian subspace estimation algorithms.

Goal 5) Novel nonlinear generalizations

We have begun study into three nonlinear models that are generalizations of the low-dimensional subspace model: Generalized Single Index Models, Variety Models, and the Union of Subspaces Model. Our specific objectives were to develop algorithms for learning these interesting nonlinear models from incomplete or undersampled measurements. Our Calibrated Single Index (CSI) algorithm for learning generalized single index models was published and presented in AAAI 2017. Our work on matrix completion with Variety Models was published in ICML 2017. Finally our work on semi-supervised subspace clustering for the union of subspaces model was also published in ICML 2017.

For Generalized SIMs, we developed the first ever algorithm to estimate a single index model in high dimensions. Our approach gave a highly flexible framework allowing many kinds of structure in the data (sparsity, group sparsity, low-rank structure for example) and arbitrary nonlinear monotonic transformations of the observations. Our algorithm outperforms other state-of-the-art approaches including simple neural networks that require a good deal more data and computation time.

For variety models, the key observation in our work studying the variety model is that despite the data matrix being high-rank, when lifted using the polynomial kernel, this data matrix is low-rank. Therefore we can perform low-rank matrix completion in the lifted space (or without lifting if we use a kernelized algorithm). This algorithm is the first tractable algorithm that can handle estimation of the variety model with missing data. Our mathematical results include a bound on the rank of general varieties, and we also identified the specific bound for the union of subspaces model, both of which were previously unknown.

Finally for the union of subspaces model, we developed a novel notion of margin for active learning of a subspace classifier, and this forms the basis of a new margin-based active learning algorithm. This algorithm was shown empirically to far outperform state-of-the-art algorithms in active classification for several computer vision datasets. For example, on the US Postal Service digits dataset, our algorithm achieves under 3% classification error with only 600 queries, whereas with 1000 queries the state-of-the-art algorithm has about 20% classification error.

Training Opportunities: Throughout the performance period of this grant, the students involved have had many professional development opportunities. This includes summer schools and workshops on topics related to the grant, poster presentations at the University of Michigan, conference attendance including poster presentations and oral presentations, and involvement with the Michigan statistical machine learning reading group. We have had undergraduate and MS students experience their first research project, and PhD students had the opportunity to mentor those younger students.

Here I also list some specific opportunities that were supported at least in part by the grant. Dejiao Zhang presented her work at AI stats and ICASSP conferences and did a summer internship at Technicolor where she studied the union of subspaces model. John Lipor attended the Simons workshop for active learning in February and presented his work at IEEE CAMSAP and ICML. David Hong and Chenlan Wang attended the Midwest Machine Learning Workshop (first annual) and presented their work there. David Hong also presented his work at several on-campus University of Michigan symposia, and did a summer internship at Sandia national labs. John Lipor and David Hong presented their work at SPARS 2017 and were involved in the co-located Pulsar Information Processing workshop. All students were involved in the Michigan Statistical Machine Learning reading group.

Results Dissemination: Our work was disseminated extensively by conference publication and presentation — Allerton conference, AAAI, ICASSP, SPARS, and ICML. We also have journal papers in submission to the Journal of Multivariate Analysis and SIAM Journal of Optimization. I also presented the active learning algorithm for union of subspaces data at the Simons workshop for Interactive Learning in February 2017. I gave several invited talks at universities, such as Johns Hopkins, Claremont McKenna, Colorado School of Mines, Toyota Technical Institute, and the University of Chicago. I also gave a seminar at the Air Force Research Lab in summer 2016. I gave a major seminar that covered most of the work from the ARO grant at the MICDE (Michigan Institute for Computational Discovery and Engineering) in April 2017. I gave a seminar specifically on our monotonic matrix completion results at the MIDAS (Michigan Institute for Data Science) seminar series in March 2017. Both these last two seminar series are interdisciplinary series that involve the entire Michigan campus.

RPPR Final Report
as of 08-Dec-2017

Honors and Awards: In 2015 PI Balzano's research partially supported by ARO earned her the Intel Early Career Faculty Fellowship.

David Hong's poster presentations won awards at the Michigan Institute of Data Science annual symposium (university wide), the Michigan Engineering Graduate Symposium (college of engineering wide), and the Michigan Student Symposium for Interdisciplinary Statistical Sciences (university wide).

Chenlan Wang won the Rackham International Student Fellowship, a fellowship awarded by the University of Michigan graduate school.

John Lipor is an NSF graduate research fellow. His travel and PI Balzano's support of his project were paid for by this ARO grant. He received a best paper honorable mention at CAMSAP 2015 for his work on active subspace clustering. John completed his dissertation in September 2017 and is continuing to a faculty position at Portland State University beginning January 2018.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Laura Balzano
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Undergraduate Student
Participant: Chenlan Wang
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Undergraduate Student
Participant: Saket Dewangan
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Dejiao Zhang
Person Months Worked: 8.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

RPPR Final Report
as of 08-Dec-2017

Participant Type: Graduate Student (research assistant)
Participant: John Lipor
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: David Hong
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE Global Conference on Signal and Information Processing
Date Received: 10-Oct-2017 Conference Date: 03-Dec-2014 Date Published: 03-Dec-2014
Conference Location: Atlanta, Georgia
Paper Title: Online Completion of Ill-conditioned Low-Rank Matrices
Authors: Ryan Kennedy, Camillo J. Taylor, Laura Balzano
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: International Conference on Machine Learning (ICLM)
Date Received: 30-Aug-2016 Conference Date: 19-Jun-2016 Date Published: 19-Jun-2016
Conference Location: New York City
Paper Title: On Learning High Dimensional Structured Single Index Models
Authors: Nikhil Rao, Ravi Ganti, Laura Balzano, RebeccaWillett, Rob Nowak
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Allerton Conference on Communications, Control, and Computing
Date Received: 10-Oct-2017 Conference Date: 27-Sep-2016 Date Published: 27-Sep-2016
Conference Location: University of Illinois at Urbana-Champaign
Paper Title: Online Sparse and Orthogonal Subspace Estimation from Partial Information
Authors: Pengyu Xiao, Laura Balzano
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 08-Dec-2017

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: AISTATS 2016: The 19th International Conference on Artificial Intelligence and Statistics
Date Received: 30-Aug-2016 Conference Date: 07-May-2016 Date Published: 07-May-2016
Conference Location: Cadiz, Spain
Paper Title: Global Convergence of a Grassmannian Gradient Descent Algorithm for Subspace Estimation
Authors: Dejiao Zhang, Larua Balzano
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 54th Allerton Conference on Communications, Control and Computing
Date Received: 10-Oct-2017 Conference Date: 28-Sep-2016 Date Published: 28-Sep-2016
Conference Location: University of Illinois at Urbana-Champaign
Paper Title: Necessary and Sufficient Conditions for Sketched Subspace Clustering
Authors: Daniel Pimentel-Alarc´on, Laura Balzano, Robert Nowak
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 54th Annual Allerton conference on Communication, Control and Computing
Date Received: 10-Oct-2017 Conference Date: 27-Sep-2017 Date Published: 27-Sep-2017
Conference Location: Monticello, Il
Paper Title: Towards a Theoretical Analysis of PCA for Heteroscedastic Data
Authors: David Hong, Laura Balzano, Jeff A. Fessler.
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2017 International Conference on Machine Learning
Date Received: 10-Oct-2017 Conference Date: 06-Aug-2017 Date Published: 13-Sep-2017
Conference Location: Sydney, Australia
Paper Title: Leveraging Union of Subspace Structure to Improve Constrained Clustering
Authors: John Lipor, Laura Balzano
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2017 International Conference on Sampling Theory and Applications
Date Received: 10-Oct-2017 Conference Date: 06-Aug-2017 Date Published: 06-Aug-2017
Conference Location: Sydney, Australia
Paper Title: Mixture Regression as Subspace Clustering
Authors: Daniel Pimentel-Alarc´on, Laura Balzano, Roummel Marcia, Robert Nowak, Rebecca Willett
Acknowledged Federal Support: **N**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: International Conference on Machine Learning
Date Received: 10-Oct-2017 Conference Date: 10-Jul-2018 Date Published:
Conference Location: Stockholm, Sweden
Paper Title: Algebraic Variety Models for High-Rank Matrix Completion
Authors: Greg Ongie, Rebecca Willett, Robert D. Nowak, Laura Balzano
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 08-Dec-2017

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing
Date Received: 10-Oct-2017 Conference Date: 05-Mar-2017 Date Published: 05-Mar-2017
Conference Location: New Orleans, Louisiana
Paper Title: MATCHED SUBSPACE DETECTION USING COMPRESSIVELY SAMPLED DATA
Authors: Dejiao Zhang, Laura Balzano
Acknowledged Federal Support: **N**

DISSERTATIONS:

Publication Type: Thesis or Dissertation
Institution: University of Michigan
Date Received: 07-Dec-2017 Completion Date: 9/9/17 2:32AM
Title: Sensing Structured Signals with Active and Ensemble Methods
Authors: John Lipor
Acknowledged Federal Support: **Y**

Subspace Methods for Massive and Messy Data
Laura Balzano, University of Michigan

Final Year Progress Report

Contents

1 Statement of the Problem Studied and Major Goals

This proposal focused on subspace estimation in various modern big-data contexts, where data are massive, streaming, time-varying, and have missing, corrupted, and ill-conditioned data. We formulated our general problem as solving for $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{T\times r}$ to be a factored representation for a data matrix $X \in \mathbb{R}^{n\times T}$. $H : \mathbb{R}^{n\times T} \to \mathbb{R}^m$ represents a linear measurement operator and $\mathcal{G}(n, r)$ is the Grassmann manifold of rank-r subspaces of $\mathbb{R}^n$, and we get the following non-convex problem:

$$\begin{aligned} \underset{U\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{T\times r}}{\text{minimize}} \quad & f(X, UV^T) = \|H\left(X - UV^T\right)\|_F^2 \\ \text{subject to} \quad & U \in \mathcal{G}(n, r) \end{aligned} \tag{1}$$

In this context we have continued working on our four main goals:

1. Algorithms and theory for estimating U when X is ill-conditioned, *i.e.,* the first and r^{th} singular values have a large ratio, or when there is a low SNR, *i.e.,* the $(r+1)$ to n^{th} singular values are relatively large compared to the first r singular values. We also extended this in two new directions: a) clusters of data with different SNR levels, and b) data that are transformed by an unknown monotonic calibration function.

2. Theory for estimating U with the natural Grassmannian incremental gradient descent algorithm when H is a compressive measurement operator.

3. Algorithms and theory for the problem variant where we minimize the ℓ_1-norm cost instead of the Frobenius norm (robust PCA) or where we add a constraint that U should be sparse (sparse PCA).

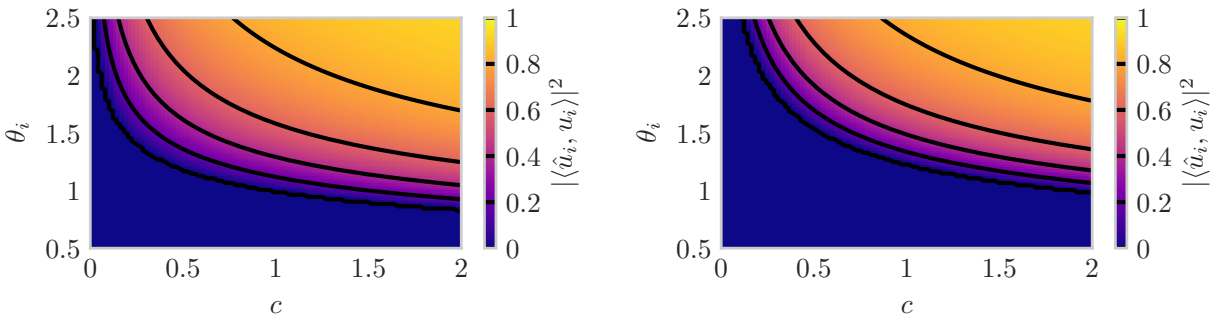
4. Applications in computer vision and Unifying convergence theory across these non-convex problems.

2 Summary of the Most Important Results

During this reporting period we have made progress on theory for estimating U when the data are heteroscedastic, and theory for estimating U from compressive measurements with the natural Grassmannian incremental gradient descent. We have also moved into studying nonlinear models more extensively after the results in Monotonic Matrix Completion from the last reporting period. These problems have provided fascinating extensions of our work on subspaces, excellent application in computer vision, and novel non-convex problem formulations of interest.

2.1 PCA estimates with heteroscedastic data

As in the last reporting period, we considered a new context wherein data from different sources have different levels of noise (*i.e.,* heteroscedastic data). Our goal in studying this problem was simply understanding how a simple PCA approach behaves in this context. It was not initially obvious how to use data from two sources with very different quality. What we learned is that, actually, sometimes it is better to throw away the data with higher noise variance, even if it is the larger quantity of data. However, when to do that depends on the relative portions of data as well as the SNR. Our initial paper was presented at Allerton in September 2016 [1]. Since that work, we have significantly simplified the expression for our predictions of the principal components, and we have extended these to predict the behavior of the estimated singular values and right singular



(a) Homoscedastic noise with $\sigma_1^2 = 1$.

(b) Heteroscedastic noise with 80% at $\sigma_1^2 = 0.8$ and 20% at $\sigma_2^2 = 1.8$.

Figure 1: Asymptotic subspace recovery of the i-th principal component as a function of sample-to-dimension ratio c and subspace amplitude θ_i with average noise variance equal to one. Contours are overlaid in black. The phase transition in (b) is further right than in (a); more samples are needed to recover the same strength signal.

vectors. We have a much deeper understanding of how the different noise variances and proportions of points will affect the outcome of PCA, and we have done an extensive comparison of our theory to classical high-dimensional PCA analysis approaches; all these new results can be found in our preprint submitted to the Journal of Multivariate Analysis [2]. These extensions are now allowing us to do new work designing a PCA weighting scheme to achieve the optimal performance.

2.2 Subspace tracking from compressive measurements

Subspace tracking Our previous work on the Grassmannian-descent subspace tracking algorithm GROUSE (Grassmannian Rank-One Update Subspace Estimation) [3] focused on subspace estimation with *missing data*. Prior work had proved local convergence [4] for Grassmannian estimation for subspace estimation with missing data. Besides global convergence results already mentioned, we also had the goal to extend this to the more general undersampling framework of compressive measurements. Since then, we have made progress on convergence results for both the case of compressive measurements as well as stronger results for missing data. This work [5] has been in the review process since September 2016. We are in the process of a major revision, and will be resubmitting soon to the SIAM journal of optimization.

Undersampled noiseless data For undersampled data, we consider two typical cases, missing data and compressively sampled data. When learning a d-dimensional subspace of R^N, under mild conditions, we can prove that with probability exceeding $1-n^{\delta d/2}$ (for some $\delta \in (0,1)$), the following unified framework holds for both cases as long as we have $m \geq O(d \log n)$ samples:

$$\mathbb{E}\left[\zeta_{t+1} \middle| U_t\right] \geq \left(1 + \eta \frac{m}{n} \frac{1-\zeta_t}{d}\right) \zeta_t \tag{2}$$

where $\eta \approx 1$ is slightly different for each sampling type [5].

Weighted Step Size Scheme for Noisy Data We also have worked on theoretical extensions to noisy data. Given a parameter α, a weighted step size scheme $\theta_t = \arctan\left((1-\alpha)\frac{\|r\|}{\|p\|}\right)$, where r is the subspace residual and p is the subspace projection, allows similar results for noisy data. We restrict $\alpha \in [0,1)$ with the goal that $\alpha \to 1$ as the range of our estimate converges to the range of the underlying true subspace; *i.e.,* $R(U_t) \to R(\bar{U})$. The intuition behind this strategy is that

dataset	SLR	SQH	SLS	CSI	Slisotron	SLNN
link ($d = 1840, n = 1051$)	0.976	0.946	0.908	**0.981**	0.959	0.975
page ($d = 3000, n = 1051$)	0.987	0.912	0.941	0.997	0.937	**0.999**
ath-rel ($d = 17785, n = 1427$)	0.857	0.726	0.733	**0.879**	0.826	0.875
aut-mot ($d = 16347, n = 1986$)	0.916	0.837	0.796	**0.941**	0.914	0.923
cryp-ele ($d = 22293, n = 1975$)	0.960	0.912	0.834	0.990	0.910	**0.994**
mac-win ($d = 7511, n = 1946$)	0.636	0.615	0.639	0.646	0.616	**0.649**

Table 1: AUC values for various methods on several datasets. The entries in bold are the best values.

the noisy part will gradually dominate $\|r\|$, we hope to include less and less information from the projection residual to our estimations as $R(U_t) \to R(\bar{U})$.

Theorem 1. *Suppose the entries of the additive noise vector ξ_t are independent and identically distributed Gaussian random variables such that $\mathbb{E}\left[\|\xi_t\|^2/\|v_t\|^2 \middle| v_t\right] \leq \sigma^2$, where v_t is the signal vector (so this lower bounds SNR). Then with probability at least $1 - n^{\delta_1 d/2}$, for some $\delta_1 \in (0, 1)$, we obtain*

$$\mathbb{E}\left[\zeta_{t+1} \middle| U_t\right] \geq \left(1 + \eta_1 \frac{m}{n} \frac{1-\zeta_t}{d} \left(1 - \frac{\sigma^2}{\frac{(1-\zeta_t)}{d} + \sigma^2}\right)\right) \zeta_t \tag{3}$$

Subspace detection and tracking from more general compressive measurements Inspired by very recent and very powerful results in high-dimensional probability, this past year we continued working on an extension of matched subspace detection with compressive measurements. The exciting contribution here is that we are able to define a clean and simple subspace detector with excellent guarantees for *any sub-gaussian sampling matrix*. This includes Bernoulli sampling, sparse matrices, missing data, any bounded matrix measurement operator, and of course Gaussian matrices as well. This work was presented at ICASSP in April 2017 and early results were described in our last report.

Using these results, we have been working on making our subspace tracking results more general. As we continue to generalize our results, we also work on demonstrating applications to more general compressive and structured measurement operators.

2.3 Novel Nonlinear Generalizations

In this our final reporting period, we worked on several exciting new nonlinear generalizations that formed the basis of our next grant proposal to ARO. While linear subspaces are useful for capturing coarse structure in data, nonlinear models have the potential to perform more detailed inference. However, nonlinear models can require much more data to estimate and how to estimate model parameters with missing data is as yet not understood. We developed some exciting new results for three nonlinear models: the generalized single index model, the algebraic variety model, and the union of subspaces model.

2.3.1 Generalized Single Index Model

In our last reporting period, we discussed new work on Monotonic Matrix Completion. This work uses the single index model to allow nonlinear modeling of low-rank matrix data. Since that time we extended the work to allow more generalized sparse structures via the atomic set model [6], including sparse, group-sparse, and low-rank models as examples. Our Calibrated Single Index (CSI) algorithm was published and presented in AAAI 2017 [7].

dataset	SLR	SQH	SLS	CSI	Slisotron	SLNN
link	0.954	0.966	0.946	**0.973**	0.914	0.939
page	0.947	0.947	0.931	**0.977**	0.925	0.962
ath-rel	0.687	0.696	0.598	0.771	0.726	**0.794**
aut-mot	0.801	0.795	0.542	**0.858**	0.812	0.855
cryp-ele	0.888	0.899	0.723	**0.949**	0.916	0.939
mac-win	0.605	0.615	0.630	**0.632**	0.579	0.590

Table 2: Classification accuracy for various methods and on several datasets. The entries in bold are the best values.

These tables compare our method with several other algorithms, in various high dimensional structural settings and on several datasets for standard sparse parameter recovery. We compare CSI with the following algorithms

- Sparse classification with the logistic loss **(SLR)** and the squared hinge loss **(SQH)**. We vary the regularization parameter over $\{2^{-10}, 2^{-9}, \cdots, 2^{9}, 2^{10}\}$. We used MATLAB code available in the L1-General library[1].
- Sparse regression using least squares **SLS**. We used a modified Frank Wolfe method [8][2], and varied the regularizer over $\{2^{-5}, 2^{-4}, \cdots, 2^{19}, 2^{20}\}$.
- Our method CSI. We varied the sparsity of the solution as $\{d/4, d/8, d/16, \cdots, d/1024\}$, rounded off to the nearest integer, where d is the dimensionality of the data.
- **Slisotron** [9] which is an algorithm for learning SIMs in low-dimensions.
- Single layer feedforward NN (**SLNN**) trained using Tensorflow [10] and the Adam optimizer [11]. We used the early stopping method and validated results over multiple epochs between 50 and 1000, and the number of hidden units were varied between 5 and 1000. The choice of single layer and not multi-layer is motivated by the fact that an SLNN would provide the fairest comparison to SIM.

We always perform a $50 - 25 - 25$ train-validation-test split of the data, and report the results on the test set. We tested the algorithms on several datasets: link and page are datasets from the UCI machine learning repository. We also use four datasets from the 20 newsgroups corpus[3]: atheism-religion, autos-motorcycle, cryptography-electronics and mac-windows. We compared the AUC in Table (1) as well as the classification accuracy in Table (2) for each of the methods. The following is a summary:

- CSI outperforms simple, widely popular learning algorithms such as SLR, SQH, SLS. Often, the difference between CSI and these other algorithms is quite substantial. For example when measuring accuracy, the difference between CSI and either SLR, SQH, SLS on all the datasets is at least 2% and in many cases as large as $4 - 5\%$.

[1] https://www.cs.ubc.ca/~schmidtm/Software/L1General.html
[2] http://www.cs.utexas.edu/~nikhilr/Code.html
[3] http://qwone.com/~jason/20Newsgroups/

- CSI comfortably outperforms Slisotron on all datasets and often by a margin as large as $5-6\%$ on accuracy scale. This is expected because Slisotron does not enforce any structure such as sparsity in its updates.
- The most interesting result is the comparison of CSI with SLNN. In spite of the simplicity of the SIM and the proposed CSI algorithm, we see that our approach is comparable to and often outperforms SLNN. This makes the CSI algorithm a valuable tool for practical data analysis.

2.3.2 Variety Model

Algebraic varieties are a direct generalization of the linear subspace and allow for very interesting nonlinear structures, some of which are illustrated in Figure 2. These nonlinear models can capture interesting nonlinearity in the data, but until our work it was unknown whether they can be recovered from partial measurements. This work was presented at ICML 2017 [12].

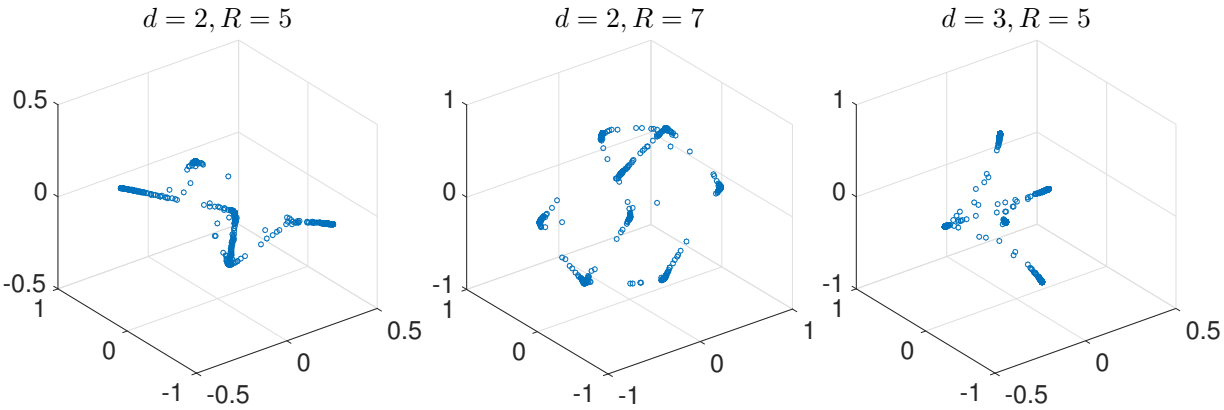


Figure 2: Data belonging to algebraic varieties in $\mathbb{R}^3$. The original data is full rank, but a nonlinear embedding of the matrix to a feature space consisting of monomials of degree at most d is low-rank with rank R, indicating the data has few degrees of freedom.

The key observation in our work studying the variety model is that despite the data matrix being high-rank, when lifted using the polynomial kernel, this data matrix is low-rank. Therefore we can perform low-rank matrix completion in the lifted space (or without lifting if we use a kernelized algorithm). We showed a bound on the rank of general varieties, and we also identified the specific bound for the union of subspaces model. If $\boldsymbol{X}$ is the original data matrix, we denote the lifted matrix using a degree d polynomial kernel as $\phi_d(\boldsymbol{X})$.

Theorem 2. *If the columns of a matrix $\boldsymbol{X}^{n\times s}$ belong to a union of k affine subspaces each of dimension at most r, then*

$$\mathit{rank}\,\phi_d(\boldsymbol{X}) \leq k\binom{r+d}{d}, \quad \textit{for all } d \geq 1. \tag{4}$$

Using this result and a back-of-the-envelope calculation, we saw that the number of samples per column m should be $m \approx O(k^{\frac{1}{d}}r)$. This approximation played out well in numerical experiments [12] in both synthetic and real computer vision data.

2.3.3 Union of Subspaces Model

We have done extensive work estimating the union of subspaces model, and our most exciting accomplishment in this reporting period was a novel semi-supervised clustering algorithm for the union of subspaces data model.

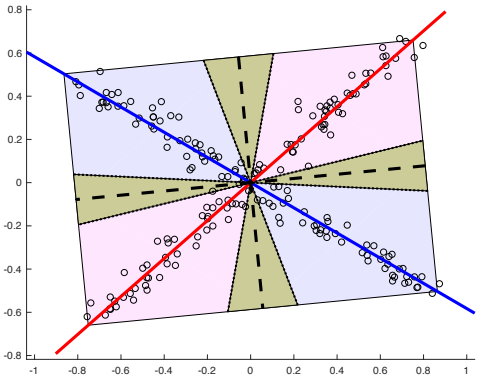


Figure 3: Illustration of subspace margin. The blue and red lines are the generative subspaces, with corresponding disjoint decision regions. The yellow-green color shows the region within some margin of the decision boundary, given by the dotted lines.

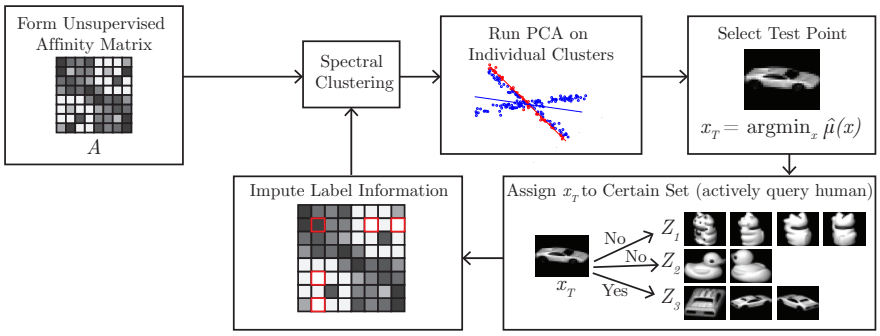


Figure 4: Diagram of SUPERPAC algorithm for pairwise constrained clustering.

In our work published and presented at ICML [13], we defined a new notion of *margin* that applies to data classified by subspace. The point of minimum margin is one equidistant to its two closest subspaces. This notion is illustrated in Figure 3, where the yellow-green color shows the region within some margin of the decision boundary. This new notion of margin allowed us to develop a new semi-supervised active learning algorithm to classify and cluster points by nearest subspace. The algorithm is illustrated in Figure 4. It actively selects a test point to query for a label, which in this case comes in the form of a pairwise comparison instead of a direct class label. Experimental results for this algorithm, which we called SUPERPAC (SUbsPace clustERing with Pairwise Active Constraints), were phenomenal. SUPERPAC significantly outperforms the state of the art algorithms. These results are showin Figure 5, where we can see that SUPERPAC uses easily less than half of the queries of the best known algorithm URASC (Uncertainty Reducing Active Spectral Clustering) to get nearly perfect clustering.

3 Future Work

Low-dimensional modeling of high-dimensional data under sensing and computational constraints constitutes a fundamental modern statistical signal processing challenge. Our work in this ARO funded research has focused on linear low-dimensional modeling with linear measurements. These models are prevalent in data science and have deep and extensive mathematical foundations that have allowed centuries of progress, and with our past three years we have made great progress on understanding linear models more deeply. However, linear models are limited in their ability to describe interesting nonlinearities in signals and measurements that are often highly relevant

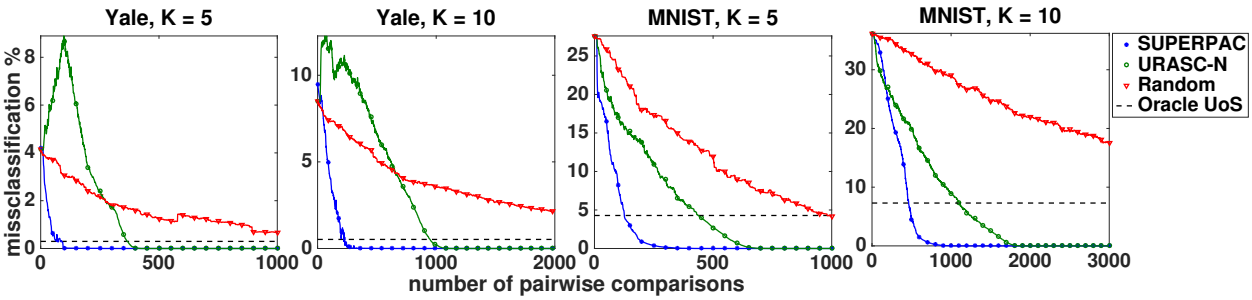


Figure 5: Misclassification rate for Yale B and MNIST datasets with many pairwise comparisons. Left-to-right: Yale B $K = 5$ (input from SSC), Yale B $K = 10$ (input from SSC), MNIST $K = 5$ (input from TSC), MNIST $K = 10$ (input from TSC).

for decision making based on collected data. On the other hand, nonlinear generalizations are often problem-specific, requiring distinct mathematical tools for addressing the technical issues of applying each nonlinear model.

We proposed a new research project to ARO that will focus on broader generalizations to nonlinear signal and measurement models, developing the foundational mathematics, algorithms, and algorithmic theory for identifying these models in high dimensions. The next research projects will focus on (1) Correlation and nonlinearities in the measurement system: measurements are often modeled as a linear operator applied to the true signal of interest, with subsampling or compression modeled as random and independent. However, many real measurement systems have nonlinearities and correlation that cannot be accurately removed a priori. (2) Algebraic variety models for relationships in the data, building off our work in Section 2.3.2. (3) Atomic set models with nonlinear measurements, building off our work in Section 2.3.1.

4 Bibliography

References

[1] D. Hong, L. Balzano, and J. Fessler. Towards a theoretical analysis of PCA for heteroscedastic data. In *Allerton Conference on Communications, Control, and Computing*, 2016.

[2] David Hong, Laura Balzano, and Jeffrey A Fessler. Asymptotic performance of PCA for high-dimensional heteroscedastic data. *arXiv preprint arXiv:1703.06610*, 2017.

[3] Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *Proceedings of the Allerton conference on Communication, Control, and Computing*, 2010.

[4] Laura Balzano and Stephen J Wright. Local convergence of an algorithm for subspace identification from partial data. *Foundations of Computational Mathematics*, pages 1–36, 2014.

[5] Dejiao Zhang and Laura Balzano. Convergence of a grassmannian gradient descent algorithm for subspace estimation from undersampled data. *Submitted to IEEE Transactions on Information Theory*, 2016. Preprint available at https://arxiv.org/abs/1610.00199.

[6] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[7] Ravi Ganti, Nikhil Rao, Laura Balzano, Rebecca Willett, and Robert Nowak. On learning high dimensional structured single index models. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[8] Nikhil Rao, Parikshit Shah, and Stephen Wright. Forward backward greedy algorithms for atomic norm regularization. *Signal Processing, IEEE Transactions on*, 63, 2015.

[9] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[10] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[11] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Greg Ongie, Rebecca Willett, Robert D Nowak, and Laura Balzano. Algebraic variety models for high-rank matrix completion. In *International Conference for Machine Learning (ICML)*, 2017. arXiv preprint arXiv:1703.09631.

[13] John Lipor and Laura Balzano. Leveraging union of subspace structure to improve constrained clustering. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.