AFRL-AFOSR-VA-TR-2019-0271

Latent Variable Graphical Modeling for High-Dimensional Data Analysis

Venkat Chandrasekaran
CALIFORNIA INSTITUTE OF TECHNOLOGY

09/12/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 12-09-2019	**2. REPORT TYPE** Final Performance	**3. DATES COVERED** *(From - To)* 15 Jun 2016 to 14 Jun 2019

4. TITLE AND SUBTITLE
Latent Variable Graphical Modeling for High-Dimensional Data Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-16-1-0210

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Venkat Chandrasekaran

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
CALIFORNIA INSTITUTE OF TECHNOLOGY
1200 E. CALIFORNIA BLDV
PASADENA, CA 91125 US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2019-0271

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

An outstanding challenge in many applications throughout science and engineering is to succinctly characterize the relationships among a large number of interacting entities. For example, in a computational biology setting a typical question involving gene regulatory networks is to discover the interaction patterns among a collection of genes in order to better understand their biological function. Similar problems also arise in analyzing word frequencies in a large corpus of text documents, in community detection in social networks, in the analysis of networks of water reservoirs in the geosciences, and in competitive interaction problems in economics. To address these challenges in a unified manner, this proposal originally aimed at developing new methodology via statistical models defined on graphs, as graphs often provide a concise representation of the interactions among a large set of variables.

Over the course of the past three years, we developed new algorithmic frameworks based on convex optimization for tasks such as associating semantics to latent variables, evaluating statistical confidence of latent variable model selection methods, finding hidden structured subgraphs inside larger networks, obtaining bounds on deviations between models specified by two networks, and fitting convex shapes to tomographic data. We demonstrate the applicability of our new methodology in domains such as reservoir modeling of the California network, hyperspectral imaging, recommender systems, and comparing molecular structure in chemistry problems.

15. SUBJECT TERMS
graphical modeling, statistical inference, latent variables, sufficient dimension reduction

16. SECURITY CLASSIFICATION OF:

a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** RIECKEN, RICHARD
Unclassified	Unclassified	Unclassified	UU		**19b. TELEPHONE NUMBER** *(Include area code)* 703-941-1100

Latent Variable Graphical Modeling for High-Dimensional Data Analysis – Final Report, AFOSR YIP

PI – Venkat Chandrasekaran; Program Manager – Doug Riecken
California Institute of Technology, Pasadena, CA 91125

September 10, 2019

Summary: An outstanding challenge in many applications throughout science and engineering is to succinctly characterize the relationships among a large number of interacting entities. For example, in a computational biology setting a typical question involving gene regulatory networks is to discover the interaction patterns among a collection of genes in order to better understand their biological function. Similar problems also arise in analyzing word frequencies in a large corpus of text documents, in community detection in social networks, in the analysis of networks of water reservoirs in the geosciences, and in competitive interaction problems in economics. To address these challenges in a unified manner, this proposal originally aimed at developing new methodology via **statistical models defined on graphs**, as graphs often provide a concise representation of the interactions among a large set of variables.

Over the course of the past three years, we developed new algorithmic frameworks based on **convex optimization** for tasks such as associating semantics to latent variables, evaluating statistical confidence of latent variable model selection methods, finding hidden structured subgraphs inside larger networks, obtaining bounds on deviations between models specified by two networks, and fitting convex shapes to tomographic data. We demonstrate the applicability of our new methodology in domains such as reservoir modeling of the California network, hyperspectral imaging, recommender systems, and comparing molecular structure in chemistry problems.

Here we describe brief outlines and outcomes of each of the projects pursued over the course of this program:

Modeling the California reservoir network [1] The recent California drought has highlighted the potential vulnerability of the state's water management infrastructure to multi-year dry intervals. Due to the high complexity of the network, dynamic storage changes in California reservoirs on a state-wide scale have previously been difficult to model using either traditional statistical or physical approaches. Indeed, although there is a significant line of research on exploring models for single (or a small number of) reservoirs, these approaches are not amenable to a system-wide modeling of the California reservoir network due to the spatial and hydrological heterogeneities of the system. In this work, we develop a state-wide statistical graphical model to characterize the dependencies among a collection of 55 major California reservoirs across the state; this model is defined with respect to a graph in which the nodes index reservoirs and the edges specify the relationships or dependencies between reservoirs. We obtain and validate this model in a data-driven manner based on reservoir volumes over the period 2003–2016. A key feature of our framework is a quantification of the effects of external phenomena that influence the entire reservoir network. We further characterize the degree to which physical factors (e.g., state-wide Palmer Drought Severity Index (PDSI), average temperature, snow pack) and economic factors (e.g., consumer price index, number of agricultural workers) explain these external influences. As a consequence of this analysis, we obtain a system-wide health diagnosis of the reservoir network as a function of PDSI.

Associating semantics to latent variables [2] Latent or unobserved phenomena pose a significant difficulty in data analysis as they induce complicated and confounding dependencies among a collection of observed variables. Factor analysis is a prominent multivariate statistical modeling approach that addresses this challenge by identifying the effects of (a small number of) latent variables on a set of observed variables. However, the latent variables in a factor model are purely mathematical objects that are derived from the observed phenomena, and they do not have any interpretation associated to them. A natural approach for attributing semantic information to the latent variables in a factor model is to obtain measurements of some additional plausibly useful covariates that may be related to the original set of observed variables, and to associate these auxiliary covariates to the latent variables. In this paper, we describe a systematic approach for identifying such associations. Our method is based on solving computationally tractable convex optimization problems, and it can be viewed as a generalization of the minimum-trace factor analysis procedure for fitting factor models via convex optimization. We analyze the theoretical consistency of our approach in a high-dimensional setting as well as its utility in practice via experimental demonstrations with real data.

Finding structured subgraphs inside larger graphs [3] Extracting structured subgraphs inside large graphs—often known as the planted subgraph problem—is a fundamental question that arises in a range of application domains. This problem is NP-hard in general and, as a result, significant efforts have been directed towards the development of tractable procedures that succeed on specific families of problem instances. We propose a new computationally efficient convex relaxation for solving the planted subgraph problem; our approach is based on tractable semidefinite descriptions of majorization inequalities on the spectrum of a symmetric matrix. This procedure is effective at finding planted subgraphs that consist of few distinct eigenvalues, and it generalizes previous convex relaxation techniques for finding planted cliques. Our analysis relies prominently on the notion of spectrally comonotone matrices, which are pairs of symmetric matrices that can be transformed to diagonal matrices with sorted diagonal entries upon conjugation by the same orthogonal matrix.

Learning convex relaxations from data [4] Regularization techniques are widely employed in optimization-based approaches for solving ill-posed inverse problems in data analysis and scientific computing. These methods are based on augmenting the objective with a penalty function, which is specified based on prior domain-specific expertise to induce a desired structure in the solution. We consider the problem of learning suitable regularization functions from data in settings in which precise domain knowledge is not directly available. Previous work under the title of 'dictionary learning' or 'sparse coding' may be viewed as learning a regularization function that can be computed via linear programming. We describe generalizations of these methods to learn regularizers that can be computed and optimized via semidefinite programming. Our framework for learning such semidefinite regularizers is based on obtaining structured factorizations of data matrices, and our algorithmic approach for computing these factorizations combines recent techniques for rank minimization problems along with an operator analog of Sinkhorn scaling. Under suitable conditions on the input data, our algorithm provides a locally linearly convergent method for identifying the correct regularizer that promotes the type of structure contained in the data. Our analysis is based on the stability properties of Operator Sinkhorn scaling and their relation to geometric aspects of determinantal varieties (in particular tangent spaces with respect to these varieties). The regularizers obtained using our framework can be employed effectively in semidefinite programming relaxations for solving inverse problems.

New relative entropy relaxations for sparse problems [5] Newton polytopes play a prominent role in the study of sparse polynomial systems, where they help formalize the idea that the root structure underlying sparse polynomials of possibly high degree ought to still be "simple." In this paper we consider sparse polynomial optimization problems, and we seek a deeper understanding of the role played by Newton polytopes in this context. Our investigation proceeds by reparametrizing polynomials as signomials – which are linear combinations of exponentials of linear functions in the decision variable – and studying the resulting signomial optimization problems. Signomial programs represent an interesting (and generally intractable) class of problems in their own right. We build on recent efforts that provide tractable relative entropy convex relaxations to obtain bounds on signomial programs. We describe several new structural results regarding these relaxations as well as a range of conditions under which they solve signomial programs exactly. The facial structure of the associated Newton polytopes plays a prominent role in our analysis. Our results have consequences in two directions, thus highlighting the utility of the signomial perspective. In one direction, signomials have no notion of "degree"; therefore, techniques developed for signomial programs depend only on the particular terms that appear in a signomial. When specialized to the context of polynomials, we obtain analysis and computational tools that only depend on the particular monomials that constitute a sparse polynomial. In the other direction, signomials represent a natural generalization of polynomials for which Newton polytopes continue to yield valuable insights. In particular, a number of invariance properties of Newton polytopes in the context of optimization are only revealed by adopting the viewpoint of signomials.

Assessing statistical confidence in latent variable models [6] Models specified by low-rank matrices are ubiquitous in contemporary applications. In many of these problem domains, the row/column space structure of a low-rank matrix carries information about some underlying phenomenon, and it is of interest in inferential settings to evaluate the extent to which the row/column spaces of an estimated low-rank matrix signify discoveries about the phenomenon. However, in contrast to variable selection, we lack a formal framework to assess true/false discoveries in low-rank estimation; in particular, the key source of difficulty is that the standard notion of a discovery is a discrete one that is ill-suited to the smooth structure underlying low-rank matrices. We address this challenge via a geometric reformulation of the concept of a discovery, which then enables a natural definition in the low-rank case. We describe and analyze a generalization of the Stability Selection method of Meinshausen and Buhlmann to control for false discoveries in low-rank estimation, and we demonstrate its utility compared to previous approaches via numerical experiments.

Fitting convex shapes to tomographic data [7] The geometric problem of estimating an unknown compact convex set from evaluations of its support function arises in a range of scientific and engineering applications. Traditional approaches typically rely on estimators that minimize the error over all possible compact convex sets; in particular, these methods do not allow for the incorporation of prior structural information about the underlying set and the resulting estimates become increasingly more complicated to describe as the number of measurements available grows. We address both of these shortcomings by describing a framework for estimating tractably specified convex sets from support function evaluations. Building on the literature in convex optimization, our approach is based on estimators that minimize the error over structured families of convex sets that are specified as linear images of concisely described sets – such as the simplex or the free spectrahedron – in a higher-dimensional space that is not much larger than the ambient space. Convex sets parametrized in this manner are significant from a computational perspective as one can optimize linear functionals over such sets efficiently; they serve a different purpose in the inferential context of the present paper, namely, that of incorporating regularization in the reconstruction while still offering considerable expressive power. We provide a geometric characterization of the asymptotic behavior of our estimators, and our analysis relies on the property that certain sets which admit semialgebraic descriptions are Vapnik-Chervonenkis (VC) classes. Our numerical experiments highlight the utility of our framework over previous approaches in settings in which the measurements available are noisy or small in number as well as those in which the underlying set to be reconstructed is non-polyhedral.

Bounding edit distance between graphs [8] The edit distance between two graphs is a widely used measure of similarity that evaluates the smallest number of vertex and edge deletions/insertions required to transform one graph to another. It is NP-hard to compute in general, and a large number of heuristics have been proposed for approximating this quantity. With few exceptions, these methods generally provide upper bounds on the edit distance between two graphs. In this paper, we propose a new family of computationally tractable convex relaxations for obtaining lower bounds on graph edit distance. These relaxations can be tailored to the structural properties of the particular graphs via convex graph invariants. Specific examples that we highlight in this paper include constraints on the graph spectrum as well as (tractable approximations of) the stability number and the maximumcut values of graphs. We prove under suitable conditions that our relaxations are tight (i.e., exactly compute the graph edit distance) when one of the graphs consists of few eigenvalues. We also validate the utility of our framework on synthetic problems as well as real applications involving molecular structure comparison problems in chemistry.

Relative entropy convex relaxations for polynomial and signomial modeling [9] We describe a generalization of the Sums-of-AM/GM Exponential (SAGE) relaxation methodology for obtaining bounds on constrained signomial and polynomial optimization problems. Our approach leverages the fact that relative entropy based SAGE certificates conveniently and transparently blend with convex duality, in a manner that Sums-of-Squares certificates do not. This more general approach not only retains key properties of ordinary SAGE relaxations (e.g. sparsity preservation), but also inspires a novel perspective-based method of solution recovery. We illustrate the utility of our methodology with a range of examples from the global optimization literature, along with a publicly available software package.

References

[1] Taeb, A., Reager, J. T., Turmon, M., and Chandrasekaran, V. (2017). A Statistical Graphical Model of the California Reservoir System. *Water Resources Research*, Vol. 53, No. 11.

[2] Taeb, A. and Chandrasekaran, V. (2018) Interpreting Latent Variables in Factor Models via Convex Optimization. *Mathematical Programming*, Vol. 167, No. 1.

[3] Candogan, U. and Chandrasekaran, V. (2018) Finding Planted Subgraphs with Few Eigenvalues using the Schur-Horn Relaxation. *SIAM Journal on Optimization*, Vol. 28, No. 1.

[4] Soh, Y. S. and Chandrasekaran, V. (2018) Learning Semidefinite Regularizers. *Foundations of Computational Mathematics*, accepted.

[5] Murray, R., Chandrasekaran, V. and Wierman, A. (2018) Newton Polytopes and Relative Entropy Optimization, *Foundations of Computational Mathematics*, under review.

[6] Taeb, A., Shah, P. and Chandrasekaran, V. (2018) False Discovery and Its Control in Low-Rank Estimation, *Journal of the Royal Statistical Society, Series B*, under review.

[7] Soh, Y. S. and Chandrasekaran, V. (2019) Fitting Tractable Convex Sets to Support Function Evaluations, *Discrete and Computational Geometry*, under review.

[8] Candogan, U. and Chandrasekaran, V. (2019) Convex Graph Invariant Relaxations for Graph Edit Distance, *Mathematical Programming*, under review.

[9] Murray, R., Chandrasekaran, V. and Wierman, A. (2019) Signomial and Polynomial Optimization via Relative Entropy and Partial Dualization, *Mathematical Programming*, under review.