AFRL-OSR-VA-TR-2013-0357

New Theory and Algorithms for Scalable Data Fusion

Martin Wainwright

UC Berkeley

JULY 2013
Final Report

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
14-07-2013	Final Report	Sep 30, 2009 -- April 30, 2013

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
New Theory and Algorithms for Scalable Data Fusion	
	5b. GRANT NUMBER
	FA9550-09-1-0466
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Martin Wainwright	
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Department of EECS, UC Berkeley, Berkeley, CA 94720 USA	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Dr. Tristan Nguyen Air Force Office of Scientific Research 875 Randolph St, Suite 325 Room 3112 Arlington, VA, 22203	AFOSR
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
	AFRL-OSR-VA-TR-2013-0357

12. DISTRIBUTION/AVAILABILITY STATEMENT

Public   DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The research performed under this grant served to address the modeling, algorithmic and theoretical challenges associated with problems of large-scale data fusion. Significant research accomplishments included: (a) the development of message passing algorithms for distributed optimization and inference; (b) the formulation and analysis of convex relaxations for estimating low-rank matrices from data; (c) the development of non-parametric methods for solving high-dimensional prediction problems; and (d) the analysis and implementation of methods for graphical model selection.

15. SUBJECT TERMS

Data fusion, scalable algorithms.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
				11	19b. TELEPHONE NUMBER (Include area code)

Reset

Final Grant Report

AFOSR grant FA9550-09-1-0466
Co-P.I. Martin Wainwright
263 Cory Hall, Department of EECS,
University of California, Berkeley
Berkeley, CA 94720

Abstract: Recent developments in sensor technology, signal processing, and communications have enabled the conception and deployment of large-scale networked sensing systems consisting of coordinated stationary and mobile platforms carrying sensors of diverse modalities. The promise of systems lies in their ability to intelligently integrate information from massive amounts of sensor data. At the core of these challenges is a fundamental information fusion task.

The research performed under this grant served to address the modeling, algorithmic and theoretical challenges associated with these problems of large-scale information fusion. Significant accomplishments include (a) the development of message-passing algorithms for distributed optimization and statistical inference, with applications to sensor fusion and computer vision; (b) the formulation and analysis of convex relaxations for estimating low-rank matrices from data, with applications to missing data problems, and tracking of dynamical systems; (c) the development of non-parametric methods for solving high-dimensional prediction problems; and (d) analysis and implementation of methods for selecting graphical models in high dimensions, with applications to terrorist cell monitoring, and social network analysis.

1 Summary

In this final report, we summarize research activity associated with AFOSR grant FA9550-09-1-0466. Overall, the grant was used to support the co-PI Prof. Martin Wainwright, as well as graduate students Sahand Negahban (Ph.D 2012, now an assistant professor at Yale University), Garvesh Raskutti (Ph. D 2012, now an assistant professor at University of Wisconsin, Madison), Alekh Agarwal (Ph.D. 2012, now research scientist at Microsoft Research), and Po-Ling Loh (Ph.D. expected in 2014). The grant has lead to the refereed conference papers [26, 24, 33, 13], as well as the journal publications [36, 2, 1, 34, 14, 37, 35, 25].

2 Honors and awards

During the grant period, Professor Martin Wainwright received an IEEE Communications Society Best Paper Award (2010), a Joint IEEE Information Theory and Communications Best Paper Award (2012), and a Medallion Lecturership from the Institute of Mathematical Statistics (2013). Graduate Student Po-Ling Loh received a Best Paper Award from the NIPS Conference in December 2012.

3 Estimation of low-rank matrices in high dimensions

In the papers [33, 34], we study the problem of estimating a matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ that is either *exactly low rank*, meaning that it has at most $r \ll \min\{p_1, p_2\}$ non-zero singular values, or more generally is *near low-rank*, meaning that it can be well-approximated by a matrix of low rank. Such exact or approximate low-rank conditions are appropriate for many applications, including multivariate or multi-task forms of regression, system identification for autoregressive processes, collaborative filtering, and matrix recovery from random projections. Analogous to the use of an ℓ_1-regularizer for enforcing sparsity, we consider the use of the nuclear norm (also known as the trace norm) for enforcing a rank constraint in the matrix setting. By definition, the nuclear norm is the sum of the singular values of a matrix, and so encourages sparsity in the vector of singular values, or equivalently for the matrix to be low-rank.

One motivation for our work is the problem of recovering system matrices in vector autogressive (VAR) processes [28]. A VAR model consists of a sequence $\{X(t)\}_{t=1}^{\infty}$, where each $X(t) \in \mathbb{R}^p$ is a vector of state variables, that evolves according to the recursion

$$X(t+1) = \Theta^* X(t) + W(t), \quad t = 1, 2, 3, \cdots,$$

where $W(t) \in \mathbb{R}^p$ are driving noise terms. Such models are widely used in different settings. They are integral parts of subspace tracking models in signal processing, motion models models in computer vision, financial data analysis, and neural data analysis (e.g., [15, 5, 10, 10]). This model and closely related ones also arise in the problem of collaborative filtering [41], in which the goal is to predict users' preferences for items (such as movies or music) based on their and other users' ratings of related items. The system matrix $\Theta^* \in \mathbb{R}^{p \times p}$ is unknown, and our goal is to estimate it, using a number of samples N *less* than the dimension of the problem. Imposing a rank r constraint on a matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ is equivalent to requiring the rows (or columns) of Θ^* lie in some r-dimensional subspace of $\mathbb{R}^{p_2}$ (or $\mathbb{R}^{p_1}$ respectively).
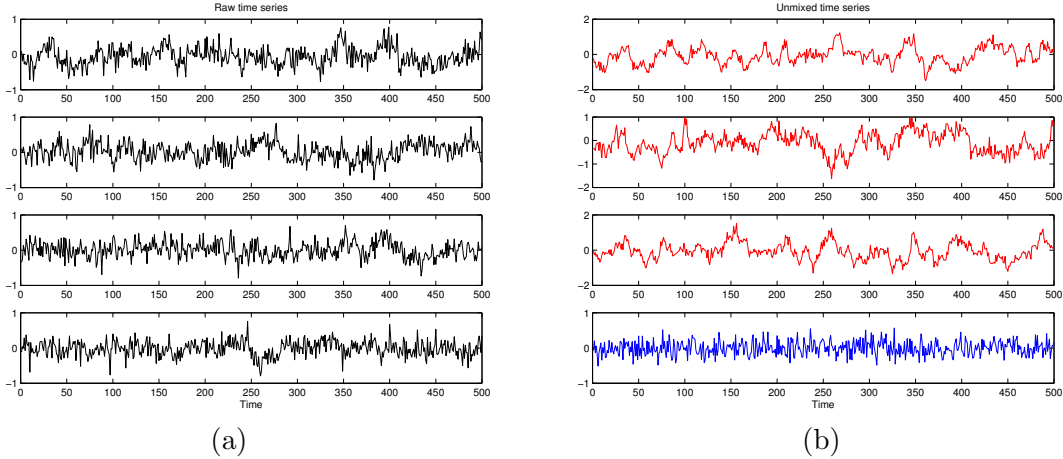
Figure 1. (a) Four entries of a $p = 100$ dimension vector autoregressive (VAR) process, generated from a system matrix $\Theta^* \in \mathbb{R}^{p \times p}$ with rank $r = 3$. Every component is a mixture of the $r = 3$ signal components with $p - r = 97$ noise components. (b) Data that has been "de-mixed" using the learned model $\widehat{\Theta}$: the first three components (in red) are estimates of the signal, whereas the remaining blue component is pure noise. Note how the signal components are much smoother than the noise component.

In order to recover low-rank matrices, we propose solving the following semidefinite program (SDP),

$$\widehat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{p_1 \times p_2}} \left\{ \frac{1}{2N} \|y - \mathfrak{X}_N(\Theta)\|_2^2 + \lambda_N \|\Theta\|_1 \right\}, \tag{1}$$

where $\lambda_N > 0$ is a regularization parameter. Here $\mathfrak{X}_N : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^N$ is an operator that maps matrices to N-vectors of observations. This optimization problem can be solved efficiently by various techniques, and we our main result is to prove upper bounds on the Frobenius norm $\|\widehat{\Theta} - \Theta^*\|_F$ between the true matrix Θ^* and the estimate $\widehat{\Theta}$. In particular, we prove that under mild conditions, the Frobenus norm error satisfies the bound

$$\|\widehat{\Theta} - \Theta^*\|_F^2 = \mathcal{O}\left(\frac{r(p_1 + p_2)}{N}\right) \tag{2}$$

with probability greater than $1 - c_1 \exp(-c_2 N \lambda_N^2)$. This bound implies that it is possible to obtain a good estimate of the matrix using far fewer than $p_1 p_2$ samples as long as the rank r is small. These theoretical results provide a remarkably good characterization of the high-dimensional scaling of this method; see Figure 2 for some illustrative results.

4   Dual Averaging for Distributed Optimization

We consider an optimization problem based on functions that are distributed over a network. More specifically, let $G = (V, E)$ be an undirected graph over the vertex set $V = \{1, 2, \ldots, n\}$ with edge set $E \subset V \times V$. Associated with each $i \in V$ is convex function $f_i : \mathbb{R}^d \to \mathbb{R}$, and our overarching goal is to solve the optimization problem

$$\min_x \frac{1}{n} \sum_{i=1}^n f_i(x) \qquad \text{such that} \ \ x \in \mathcal{X}, \tag{3}$$
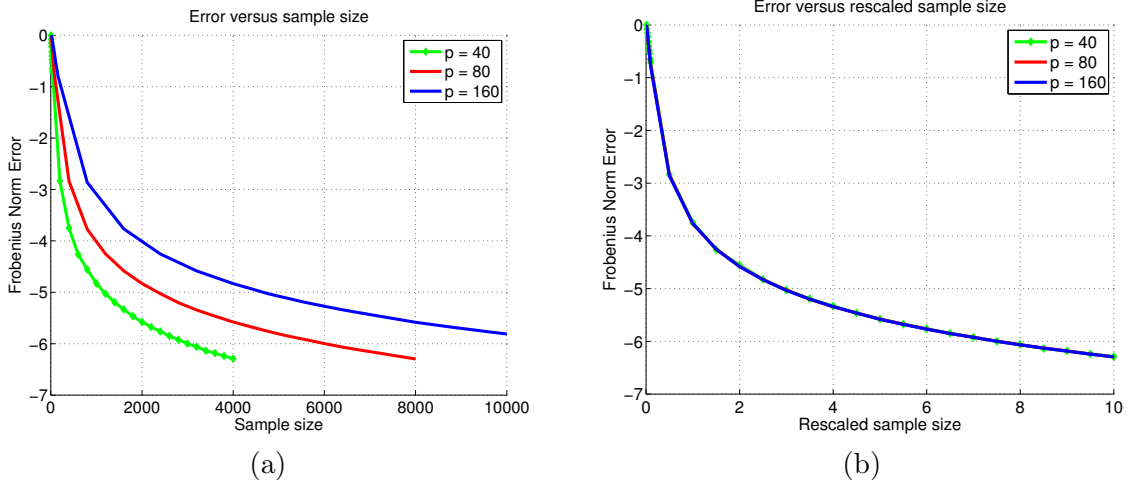
Figure 2. Results of applying the SDP (1) with nuclear norm regularization to the problem of low-rank multivariate regression. (a) Plots of the Frobenius error $\|\widehat{\Theta} - \Theta^*\|_F$ on a logarithmic scale versus the sample size N for three different matrix sizes $p \in \{40, 80, 160\}$, all with rank $r = 10$. (b) Plots of the same Frobenius error versus the rescaled sample size $N/(rp)$. Consistent with theory, all three plots are now extremely well-aligned.

where $\mathcal{X}$ is a closed convex set. Each function f_i is convex and hence sub-differentiable, but need not be smooth. We assume without loss of generality that $0 \in \mathcal{X}$, since we can simply translate $\mathcal{X}$. Each node $i \in V$ is associated with a separate agent, and each agent i maintains its own parameter vector $x_i \in \mathbb{R}^d$. The graph G imposes communication constraints on the agents: in particular, agent i has local access to only the objective function f_i and can communicate directly only with its immediate neighbors $j \in \mathcal{N}(i) := \{j \in V \mid (i, j) \in E\}$.
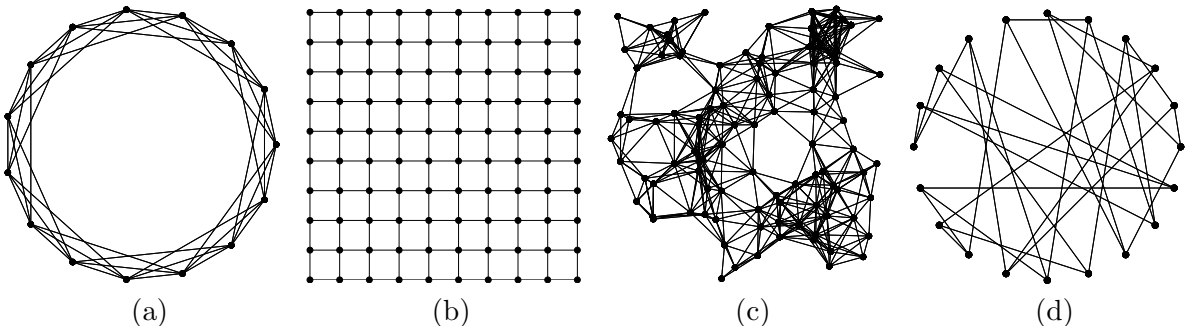

Figure 3. Illustration of some graph classes of interest in distributed protocols. (a) A 3-connected cycle. (b) Two-dimensional grid with 4-connectivity, and non-toroidal boundary conditions. (c) A random geometric graph. (d) A random 3-regular expander graph.

Problems of this nature arise in a variety of application domains, and as motivation for the analysis to follow, let us consider a few here. A first example is a sensor network, in which each agent represents a sensor mote, equipped with a radio transmitter for communication, some basic sensing devices, and some local memory and computational power. In environmental applications of sensor networks, each mote i might take a measurement y_i of the temperature, and the global objective could be to compute the median of the measurements $\{y_1, y_2, \ldots, y_n\}$. This median computation problem can be formulated as minimizing

4

the scalar objective function $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$, where $f_i(x) = |x - y_i|$. Similar formulations apply to the problem of computing other statistics such as means, variances, quantiles and other M-estimators. Another motivating example is the machine learning problem of classification, in which $\mathcal{X}$ is the parameter space of the statistician or learner. Each function f_i is the empirical loss over the subset of data assigned to the ith processor, and assuming that each subset is of equal size (or that the f_i are normalized suitably), the average f is the empirical loss over the entire dataset. Here we use cluster computing as our computational model, where each processor is a node in the cluster, and the graph G contains edges between those processors that are directly connected with small network latencies.

We consider an appropriate and novel extension of dual averaging to the distributed setting. At each iteration $t = 1, 2, 3, \ldots$, the algorithm maintains n pairs of vectors $(x_i(t), z_i(t)) \in \mathcal{X} \times \mathbb{R}^d$, with the ith pair associated with node $i \in V$. At iteration t, each node $i \in V$ computes an element $g_i(t) \in \partial f_i(x_i(t))$ in the subdifferential of the local function f_i and receives information about the parameters $\{z_j(t),\ j \in \mathcal{N}(i)\}$ associated with nodes j in its neighborhood $\mathcal{N}(i)$. Its update of the current estimated solution $x_i(t)$ is based on a convex combination of these parameters. To model this weighting process, let $P \in \mathbb{R}^{n \times n}$ be a symmetric matrix of non-negative weights that respects the structure of the graph G, meaning that for $i \neq j$, $P_{ij} > 0$ only if $(i, j) \in E$. We assume that P is a doubly stochastic matrix, so that

$$\sum_{j=1}^{n} P_{ij} = \sum_{j \in \mathcal{N}(i)} P_{ij} = 1 \quad \text{for all } i \in V, \quad \text{and} \quad \sum_{i=1}^{n} P_{ij} = \sum_{i \in \mathcal{N}(j)} P_{ij} = 1 \quad \text{for all } j \in V.$$

Using this notation, given the non-increasing sequence $\{\alpha(t)\}_{t=0}^{\infty}$ of positive stepsizes, each node $i \in V = \{1, 2, \ldots, n\}$ performs the updates

$$z_i(t + 1) = \sum_{j \in \mathcal{N}(i)} P_{ij} z_j(t) - g_i(t), \quad \text{and} \tag{4a}$$

$$x_i(t + 1) = \Pi_{\mathcal{X}}^{\psi}(-z_i(t + 1), \alpha(t)), \tag{4b}$$

where the projection $\Pi_{\mathcal{X}}^{\psi}$ is orthogonal projection onto $\mathcal{X}$. Note that each node obtains its new dual parameter $z_i(t + 1)$ from a weighted average of its own subgradient $g_i(t)$ and the parameters $\{z_j(t), j \in \mathcal{N}(i)\}$ in its own neighborhood $\mathcal{N}(i)$, and then computes the next local iterate $x_i(t + 1)$ by a projection defined by the proximal function ψ and stepsize $\alpha(t) > 0$.

We consider several settings for distributed minimization. We study fixed communication protocols, which are of interest in a variety of areas such as cluster computing or sensor networks with a fixed hardware-dependent protocol. We also investigate randomized communication protocols as well as randomized network failures, which are often essential to handle gracefully in wireless sensor networks and large clusters with potential node failures. Randomized communication also provides interesting tradeoffs between communication savings and convergence rates. In this setting, we obtain much sharper results than previous work by studying the spectral properties of the expected transition matrix of a random walk on the underlying graph. We also present a relatively straightforward extension of our analysis for stochastic gradient information.

We provide sharp bounds on their convergence rates as a function of the network size and topology. Our analysis clearly separates the convergence of the optimization algorithm itself from the effects of communication constraints arising from the network structure. We show that the number of iterations required by our algorithm scales inversely in the spectral gap of the network. The sharpness of this prediction is confirmed both by theoretical lower bounds and simulations for various networks.

5 Sparse non-parametric regression in high dimensions

Sparsity is an attractive assumption for both practical and theoretical reasons: it leads to more interpretable models, reduces computational cost, and allows for model identifiability even under high-dimensional scaling, where the dimension p exceeds the sample size N. While a large body of work has focused on sparse linear models, many applications call for the additional flexibility provided by non-parametric models. In the general setting, a non-parametric regression model takes the form $y = f^*(x_1, \ldots, x_p) + w$, where $f^* : \mathbb{R}^p \to \mathbb{R}$ is the unknown regression function, and w is scalar observation noise. Unfortunately, this general non-parametric model is known to suffer severely from the so-called "curse of dimensionality", in that for most natural function classes (e.g., twice differentiable functions), the sample size N required to achieve any given error grows exponentially in the dimension p. Given this curse of dimensionality, it is essential to further constrain the complexity of possible functions f^*. One attractive candidate is the class of *additive non-parametric models* [17], in which the function f^* has an additive decomposition of the form

$$f^*(x_1, x_2, \ldots, x_p) = \sum_{j=1}^{p} f_j^*(x_j), \tag{5}$$

where each component function f_j^* is univariate. Given this additive form, this function class no longer suffers from the exponential explosion in sample size of the general non-parametric model. Nonetheless, one still requires a sample size $N \gg p$ for consistent estimation; note that this is true even for the linear model, which is a special case of equation (5).

A natural extension of sparse linear models is the class of *sparse additive models*, in which the unknown regression function is assumed to have a decomposition of the form

$$f^*(x_1, x_2 \ldots, x_p) = \sum_{j \in S} f_j^*(x_j), \tag{6}$$

where $S \subseteq \{1, 2, \ldots, p\}$ is some unknown subset of cardinality $|S| = s$. Of primary interest is the case when the decomposition is genuinely sparse, so that $s \ll p$. To the best of our knowledge, this model class was first introduced by [21], and has since been studied by various researchers [20, 29, 38, 45]. Note that the sparse additive model (6) is a natural generalization of the sparse linear model, to which it reduces when each univariate function is constrained to be linear.

In past work, several groups have proposed computationally efficient methods for estimating sparse additive models (6). Just as ℓ_1-based relaxations such as the Lasso have desirable properties for sparse parametric models, more general ℓ_1-based approaches have proven to be successful in this setting. [21] proposed the COSSO method, which extends the Lasso to cases where the component functions f_j^* lie in a reproducing kernel Hilbert space (RKHS); see also [45] for a similar extension of the non-negative garrote [8]. [6] analyzes a closely related method for the RKHS setting, in which least-squares loss is penalized by an ℓ_1-sum of Hilbert norms, and establishes consistency results in the classical (fixed p) setting. Other related ℓ_1-based methods have been proposed in independent work by [19], [38] and [29], and analyzed under high-dimensional scaling ($p \gg N$). Each of the above papers establish consistency and convergence rates for the prediction error under certain conditions on the covariates as well as the sparsity s and dimension p. However, it is not clear whether the rates obtained in these papers are sharp for the given methods, nor whether the rates are minimax-optimal. Past

work by [20] establishes rates for sparse additive models with an additional global boundedness condition, but as will be discussed at more length in the sequel, these rates are not minimax optimal in general.

This paper makes three main contributions to this line of research. Our first contribution is to analyze a simple polynomial-time method for estimating sparse additive models and provide upper bounds on the error in the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms. The estimator[1] that we analyze is based on a combination of least-squares loss with two ℓ_1-based sparsity penalty terms, one corresponding to an $\ell_1/L^2(\mathbb{P}_n)$ norm and the other an $\ell_1/\|\cdot\|_{\mathcal{H}}$ norm. Our first main result shows that with high probability, if we assume the univariate functions are bounded and independent, the error of our procedure in the squared $L^2(\mathbb{P}_n)$ and $L^2(\mathbb{P})$ norms is bounded by $\mathcal{O}\big(\frac{s\log p}{N} + s\delta_n^2\big)$, where the quantity δ_n^2 corresponds to the optimal rate for estimating a single univariate function. Importantly, our analysis does *not* require a global boundedness condition on the class $\mathcal{F}_s$ of all s-sparse models, an assumption that is often imposed in classical nonparametric analysis. Indeed, as we discuss below, when such a condition is imposed, then significantly faster rates of estimation are possible. The proof involves a combination of techniques for analyzing M-estimators with decomposable regularizers [32] combined with various techniques in empirical process theory for analyzing kernel classes [7, 31, 42]. Our second contribution is complementary in nature, in that it establishes algorithm-independent minimax lower bounds on $L^2(\mathbb{P})$ error. These minimax lower bounds are specified in terms of the metric entropy of the underlying univariate function classes. For both finite-rank kernel classes and Sobolev-type classes, these lower bounds match our achievable results up to constant factors in the regime of sub-linear sparsity ($s = o(p)$). Thus, for these function classes, we have a sharp characterization of the associated minimax rates. The lower bounds derived in this paper initially appeared in the Proceedings of the NIPS Conference (December 2009). The proofs of Theorem 2 is based on characterizing the packing entropies of the class of sparse additive models, combined with classical information theoretic techniques involving Fano's inequality and variants [16, 43, 44].

# 6 High-dimensional inference with graphical models

Inference methods based on graphical models are now prevalent in many fields, running the gamut from computer vision and civil engineering to political science and epidemiology. In many applications, learning the edge structure of the underlying graphical model is relevant to the researcher—for instance, a graphical model may be used to represent friendships between people in a social network, or links between organisms with the propensity to spread an infectious disease. It is well-known that zeros in the inverse covariance matrix of a multivariate Gaussian distribution indicate the absence of an edge in the corresponding graphical model. This fact, combined with techniques in high-dimensional statistical inference, has been leveraged by many authors to recover the structure of a Gaussian graphical model when the edge set is sparse (e.g., see the papers [39, 46, 11, 30] and references therein). Recently, Liu et al. [23, 22] introduced the notion of a nonparanormal distribution, which generalizes the Gaussian distribution by allowing for univariate monotonic transformations, and argued that the same structural properties of the inverse covariance matrix carry over to the nonparanormal.

However, the question of whether there exists a relationship between conditional independence and the structure of the inverse covariance matrix in a general graph remains unresolved. In this paper, we focus on discrete graphical models, and establish a number of interesting

---

[1]The same estimator was proposed concurrently by [20]; we discuss connections to this work in the sequel.

links between covariance matrices and the edge structure of the underlying graph. We show that, instead of only analyzing the standard covariance matrix, it is often fruitful to augment the usual covariance matrix with higher-order interaction terms. Our main result has a striking corollary in the context of tree-structured graphs: for *any* discrete graphical model, the inverse of a generalized covariance matrix is always (block) graph-structured. In particular, for binary variables, the inverse of the usual covariance matrix reflects the zero-pattern of the tree. We also establish a number of more general results that apply to non-tree-structured graphs, and derive several corollaries that guarantee consistency of known methods for graph selection and lead to a new method for neighborhood selection in an arbitrary sparse graph. Our methods handle noisy or missing data in a seamless manner.

Other related work on graphical model selection for discrete graphs includes the classic Chow-Liu algorithm for trees [12], nodewise logistic regression for discrete models with pairwise interactions [40, 18], and techniques based on conditional entropy or mutual information [9, 4]. Our main contribution is to present a clean and surprising result on a simple link between the inverse covariance matrix and edge structure of a discrete model, which may be used to derive inference algorithms applicable even to data with systematic corruptions.
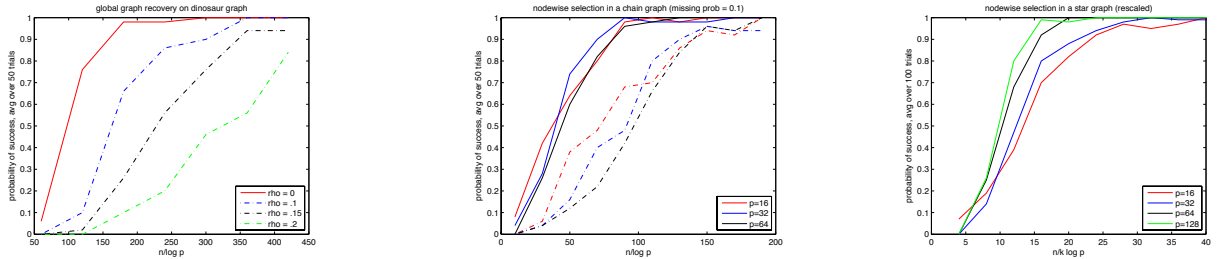
(a) Dino graph with missing data   (b) Chain graph with missing data   (c) Star graph

Figure 4. Simulation results for global and nodewise recovery methods on binary Ising models. Panel (a) shows simulation results for the log-determinant method applied to the dinosaur graph, averaged over 50 trials. Panel (b) shows simulation results for nodewise regression applied to chain graphs with different numbers of nodes, averaged over 50 trials. Panel (c) shows simulation results for nodewise regression applied to star graphs with maximal node degree $\sqrt{p}$, averaged over 100 trials. The horizontal axis gives the rescaled sample size $\frac{n}{d \log p}$.

Figure 4 depicts the results of several simulations we performed to test our theoretical predictions. In all cases, we generated binary Ising models with node weights 0.1 and edge weights 0.3 (using spin $\{-1, 1\}$ variables). Panel (a) shows the results of our global recovery method applied to the dinosaur graph. The solid curve shows the probability of success in recovering the 15 edges of the graph, as a function of the rescaled sample size $\frac{n}{\log p}$, where $p = 13$. The dotted curves show the corresponding success probabilities for missing data fractions $\rho \in \{0.1, 0.15, 0.2\}$, using the corrected estimator. We observe that all four runs display a transition from success probability 0 to success probability 1 in roughly the same range, as predicted by our theory. Indeed, since the dinosaur graph has only singleton separators, our theory ensures that the inverse covariance matrix is exactly graph-structured. Note that the curves shift right as the fraction ρ of missing data increases, since the problem becomes harder.

Panels (b) and (c) show simulation results for our nodewise regression method on chain and star graphs, with increasing numbers of nodes $p \in \{16, 32, 64\}$. The modified Lasso program was optimized using a form of composite gradient descent due to Agarwal et al. [3],

guaranteed to converge to a small neighborhood of the optimum even when the problem is non-convex [27]. Our theory predicts that, when plotted against *rescaled sample size* $\frac{n}{\log p}$, the curves for different problems should be roughly aligned, which we observe for both fully-observed samples (solid lines) and missing data (dotted lines). Again, the curves for missing data ($\rho = 0.1$) are shifted right relative to the curves for fully-observed data, since the recovery problem becomes harder with a higher fraction of missing data. Panel (c) shows similar curves for fully-observed samples from a star graph, where the central hub has degree $\sqrt{p}$. Here, we use the rescaled sample size $\frac{n}{d \log p}$, since the degree is growing.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452–2482, 2012.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 40(2):1171–1197, 2012.

[3] A. Agarwal, S. Negahban, and M.J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. Technical report, UC Berkeley, January 2012. Available at http://arxiv.org/abs/1104.4824.

[4] A. Anandkumar, V.Y.F. Tan, and A.S. Willsky. High-dimensional structure learning of Ising models : Local separation criterion. *Preprint*, June 2011.

[5] C. W. Anderson, E. A. Stolz, and S. Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. *IEEE Trans. on bio-medical engineering*, 45(3):277, 1998.

[6] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[7] P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 2005. To appear.

[8] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, (37):373–384, 1995.

[9] G. Bresler, E. Mossel, and A. Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *APPROX-RANDOM*, pages 343–356, 2008.

[10] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7(5), May 2004.

[11] T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

[12] C.I. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

[13] J. Duchi, A. Agarwal, and M. J. Wainwright. Disdain: Dual averaging for in-networking averaging. In *NIPS Conference*, Vancouver, Canada, December 2009.

[14] J. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, March 2012.

[15] L. Harrison, W. D. Penny, and K. Friston. Multivariate autoregressive modeling of fmri time series. *NeuroImage*, 19:1477–1491, 2003.

[16] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:794–798, 1978.

[17] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

[18] A. Jalali, P.D. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. *Journal of Machine Learning Research - Proceedings Track*, 15:378–387, 2011.

[19] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of COLT*, 2008.

[20] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38:3660–3695, 2010.

[21] Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34:2272–2297, 2006.

[22] H. Liu, F. Han, M. Yuan, J.D. Lafferty, and L.A. Wasserman. High dimensional semiparametric Gaussian copula graphical models. *ArXiv e-prints*, March 2012. Available at http://arxiv.org/abs/1202.2169.

[23] H. Liu, J.D. Lafferty, and L.A. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

[24] P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *NIPS Conference*, December 2011.

[25] P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, September 2012.

[26] P. Loh and M. J. Wainwright. No voodoo here! learning discrete graphical models via inverse covariance estimation. In *Neural Information Processing Systems (NIPS)*, Lake Tahoe, CA, December 2012.

[27] P. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, To appear. Available at http://arxiv.org/abs/1109.3714.

[28] H. Lütkepolhl. *New introduction to multiple time series analysis*. Springer, New York, 2006.

[29] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821, 2009.

[30] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[31] S. Mendelson. Geometric parameters of kernel machines. In *Proceedings of COLT*, pages 29–43, 2002.

[32] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012.

[33] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. In *Proceedings of the ICML Conference*, Haifa, Israel, June 2010.

[34] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[35] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, May 2012.

[36] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010.

[37] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 12:389–427, March 2012.

[38] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: sparse additive models. Technical Report arXiv:0711.4555v2, Carnegie Mellon University, 2008.

[39] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.

[40] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38:1287, 2010.

[41] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2004.

[42] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[43] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[44] B. Yu. Assouad, Fano and Le Cam. *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pages 423–435, 1996.

[45] M. Yuan. Nonnegative garrote component selection in functional ANOVA models. In *Conference on Artificial Intelligence and Statistics*, pages 660–666, 2007.

[46] M. Yuan. High-dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 99:2261–2286, August 2010.