Learning High-Dimensional DAGs: Provable Statistical Guarantees and Scalable Approximation

Bryon Aragam[†,‡] **Jiaying Gu**[†] **Arash A. Amini**[†] **Qing Zhou**[†]
[†]UCLA [‡]Carnegie Mellon University

Abstract

We introduce a recently developed score-based framework for structure learning of directed acyclic graphs (DAGs) on high-dimensional data. Compared to undirected graphs—which are well understood and for which there are algorithms that scale to millions of nodes—the situation for directed graphs is far less advanced, with methods still struggling to handle datasets with thousands of variables on commodity hardware. To address this, we developed a novel framework for DAG learning that simultaneously provides high-dimensional statistical guarantees, scalable computation to tens of thousands of nodes, and user-friendly software, in addition to being able to learn causal networks in the presence of experimental data. Furthermore, this framework avoids commonly used but uncheckable assumptions found in the literature such as faithfulness and irrepresentability, giving a sense of what happens when score-based methods are naïvely applied to high-dimensional datasets. In particular, our results yield—for the first time—finite-sample guarantees for structure learning of Gaussian DAGs in high-dimensions via score-based estimation.

1 Introduction

Despite the popularity of high-dimensional graphical models, there is a dearth of fast algorithms and guarantees for DAG learning in high-dimensions. Compared to their simpler undirected counterparts, DAGs present new statistical and computational challenges. These challenges include:

- *Nonconvexity.* The acyclicity constraint imposes a combinatorial, nonconvex constraint on the learning problem.
- *Nonsmoothness.* The acyclicity constraint also imposes a nonsmooth constraint, and this is compounded by the use of nonsmooth regularizers in high-dimensions.
- *Nonidentifiability.* Bayesian network models are not unique, and in general each permutation of the variables results in a different, minimal Bayesian network.
- *Nonpolynomial complexity.* Unlike learning undirected graphs, DAG learning is NP-hard [6]. Furthermore, reduction to neighbourhood regression involves solving a superexponential number of regression problems, compared to linear for undirected graphs.

These challenges are exacerbated in pursuing score-based learning [9], even though it is well-known to outperform other strategies [1] such as constraint-based learning [10].

In this work, we discuss our recent efforts towards understanding the statistical properties of score-based learning on high-dimensional data with $p \gg n$. Specifically, we discuss the following topics:

- *Theory.* Our recent work [2] provides the first ever structure learning guarantees for score-based learning in high-dimensions, and leverages a novel generalized neighbourhood regression framework that is of independent interest.
- *Computation.* Our recent efforts towards designing scalable approximate algorithms for learning score-based estimators from data [1, 8]. These algorithms are based on an efficient block coordinate descent scheme that scales to tens of thousands of variables.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

- *Causal inference.* Given experimental data with interventions, our framework can learn causal networks [7, 8], which allow for an intuitive, causal interpretation of DAG models.
- *Software.* Finally, we recently released a software library `sparsebn` [3], which gives end users push-button access to these methods, making it easy for practitioners to leverage our methodology on real data.

This provides a unified theoretical, computational and algorithmic framework for learning Bayesian networks from high-dimensional data. Recent work has covered some special cases [14, 11], however, these works fail to address the most fundamental problem of structure learning and do not propose tractable algorithms. To the best of our knowledge, this is the only such framework for score-based learning that simultaneously covers datasets with $p \gg n$ and p in the tens of thousands.

2 Overview

Let $X = (X_1, \ldots, X_p)$. Our approach is based on the well-known structural equation model (SEM) interpretation of DAG models. In this approach, we start by directly modeling each conditional probability distribution $\mathbb{P}(X_j \,|\, \text{pa}(X_j); \theta_j)$ via a generalized linear model. This yields a well-defined likelihood, which will be employed in the score function defined below.

Assume that the graph is parametrized by a $p \times p$ weighted adjacency matrix $B = (\beta_{kj})_{k,j=1}^p$ and let $\mathbb{D}_p$ denote the space of $p \times p$ weighted adjacency matrices that correspond to acyclic graphs. In score-based learning, one defines a score function Q and attempts to solve the following program:

$$\widehat{B} \in \operatorname*{arg\,min}_{B \in \mathbb{D}_p} Q(B). \tag{1}$$

We define $Q(B) = \ell(B) + \rho_\lambda(B)$, where ℓ is a loss function (e.g. negative log-likelihood or least squares) and ρ_λ is a regularizer such as the ℓ_1 norm or the group ℓ_2 norm. The present work covers Gaussian [1, 2] and discrete [8] models, although extensions to other probability models is natural.

2.1 Theory

The program (1) is a nonconvex, nonsmooth program that is very difficult to study. Moreover, due to the nonidentifiability of DAG models [§2.1, 2], even defining a proper notion of consistency is nontrivial. Owing to these challenges, until recently very little was known about the statistical and computational properties of these models on finite-samples, even for simple Gaussian models. Chickering [5] provided an asymptotic analysis of score-based learning under the restrictive *faithfulness* assumption, however, this analysis has not been extended to the high-dimensional setting. This should be contrasted with undirected graphical models, for which there has been a flurry of work and positive results, starting with [12].

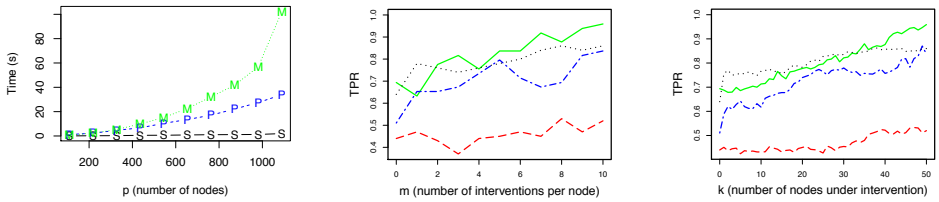
Our work [2] provides the first ever high-dimensional guarantees for $\widehat{B}$ under least-squares loss, including the following:

- Finite-sample structure learning guarantees;
- ℓ_2 rates of convergence for parameter estimation;
- Oracle inequalities;
- Upper bounds on the sparsity of $\widehat{B}$.

These results do not assume faithfulness, and by leveraging nonconvex regularizers such as the MCP, we are able to avoid restrictive *incoherence* and *irrepresentability* assumptions on the data matrix. In proving these results, we have developed a novel neighbourhood regression analysis that provides uniform, finite-sample guarantees over a family of penalized least squares estimators whose size grows as $O(p!)$. The key step is a monotonicity argument that reduces this superexponential class to a more tractable polynomial class of size $O(\text{poly}(p))$ under a sparsity assumption. In addition, these results can be applied to learning causal DAGs and conditional independence relations.

2.2 Computation

In addition to being difficult to analyze theoretically, the estimator $\widehat{B}$ is challenging to compute. Exact algorithms exist for small problems with p in the hundreds [13], however, our main interest is



(A) (B) (C)

Figure 1: (A) Timing comparison (in seconds). (solid black line) S = `sparsebn` method, (dashed blue line) P = PC algorithm, (dotted green line) M = MMHC algorithm. (B) Improvements when adding more interventions per node to four types of simulated networks (scale-free, solid green line; small-world, dashed red line; polytree, dashed blue line; bipartite graph, dashed black line). (C) Improvements when increasing the number of nodes under intervention.

in problems where p could very well be in the tens of thousands (e.g. genomics and medicine). In this regime, exact algorithms are infeasible. In order to scale to large-scale problems, we propose an approximate algorithm based on block coordinate descent:

1. Repeat outer loop until stopping criterion met:
2. *Outer loop.* For each pair (j, k), $j \neq k$:
   (a) Minimize (1) with respect to (β_{kj}, β_{jk}), holding all other parameters fixed;
   (b) If the edge $k \to j$ (resp. $j \to k$) induces a cycle in the graph, set $\beta_{kj} \leftarrow 0$ (resp. $\beta_{jk} \leftarrow 0$) and then update β_{jk} (resp. β_{kj});
   (c) Repeat inner loop until convergence:
      3. *Inner loop.* Fix the edge set E from the outer loop and minimize (1) by cycling through the edge weights β_{kj} for $(k, j) \in E$.

By avoiding traditional greedy approaches, this algorithm is extremely efficient on high-dimensional datasets for both continuous and discrete data [1, 8], and outperforms existing methods such as greedy search (GES), max-min hill climbing (MMHC), and the PC algorithm (Figure 1(A)).

2.3 Causal inference

The injection of causal information has increasingly been acknowledged as a key ingredient in modern machine learning applications [4]. DAGs are a popular representation of causal knowledge, and can be learned from experimental interventions. To see how this can be accomplished, let $\mathcal{M} \subset \{1, \ldots, p\}$ be the set of variables under intervention, so the joint probability decomposes as

$$\mathbb{P}(X_1, \ldots, X_p) = \prod_{i \notin \mathcal{M}} \mathbb{P}(X_i \,|\, \text{pa}(X_i)) \prod_{i \in \mathcal{M}} \mathbb{P}(X_i \,|\, \bullet), \tag{2}$$

where $\mathbb{P}(X_i \,|\, \bullet)$ is the marginal distribution of X_i from which experimental samples are drawn. Thus, experimental data sets generated from the true DAG $\mathcal{G}$ can be considered as data sets generated from a DAG $\mathcal{G}'$, where $\mathcal{G}'$ is obtained by removing all directed edges in $\mathcal{G}$ pointing to the variables under intervention. By leveraging the decomposition (2) in the likelihood, we are able to incorporate experimental data (even when mixed with observational data) into the learning step. Figures 1(B-C) illustrate the improvements when learning causal relationships under interventional data.

2.4 Software

Finally, we have developed the open-source `sparsebn` library for learning Bayesian networks [3].[1] This is an R package that implements the methods discussed in the previous sections, with functions for learning continuous and discrete networks in the presence of mixed observational and experimental data. This allows practitioners to accurately learn large causal networks such as genetic networks, which have important applications in understanding the genetic basis of disease.

[1] CRAN: `https://cran.r-project.org/package=sparsebn`, Source code: `https://github.com/itsrainingdata/sparsebn`.

3 Discussion

We have proposed a novel framework for learning DAGs in high-dimensions, which has important applications in genomics, medicine, and computational biology. Our results have implications for learning causal relationships in machine learning systems, as well as providing a powerful theoretical framework for analyzing graphical models. One of the most useful contributions is a framework for non-Gaussian models via generalized linear models, which is an interesting direction for future work.

4 Acknowledgements

This work was supported by NSF grant IIS-1546098.

References

[1] Aragam, Bryon, & Zhou, Qing. 2015. Concave Penalized Estimation of Sparse Gaussian Bayesian Networks. *Journal of Machine Learning Research*, **16**, 2273–2328.

[2] Aragam, Bryon, Amini, Arash A., & Zhou, Qing. 2016. Learning directed acyclic graphs with penalized neighbourhood regression. *Submitted*, **arXiv:1511.08963**.

[3] Aragam, Bryon, Gu, Jiaying, & Zhou, Qing. 2017. Learning Large-Scale Bayesian Networks with the sparsebn Package. *Submitted*, **arXiv:1703.04025**.

[4] Bottou, Léon, Peters, Jonas, Quiñonero-Candela, Joaquin, Charles, Denis X, Chickering, D Max, Portugaly, Elon, Ray, Dipankar, Simard, Patrice, & Snelson, Ed. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, **14**(1), 3207–3260.

[5] Chickering, David Maxwell. 2003. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, **3**, 507–554.

[6] Chickering, David Maxwell, Heckerman, David, & Meek, Christopher. 2004. Large-sample learning of Bayesian networks is NP-hard. *The Journal of Machine Learning Research*, **5**, 1287–1330.

[7] Fu, Fei, & Zhou, Qing. 2013. Learning Sparse Causal Gaussian Networks With Experimental Intervention: Regularization and Coordinate Descent. *Journal of the American Statistical Association*, **108**(501), 288–300.

[8] Gu, Jiayang, Fu, Fei, & Zhou, Qing. 2016. Penalized Estimation of Directed Acyclic Graphs From Discrete Data. *Submitted*, **arXiv:1403.2310**.

[9] Heckerman, David, Geiger, Dan, & Chickering, David M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, **20**(3), 197–243.

[10] Kalisch, Markus, & Bühlmann, Peter. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, **8**, 613–636.

[11] Loh, Po-Ling, & Bühlmann, Peter. 2014. High-Dimensional Learning of Linear Causal Networks via Inverse Covariance Estimation. *Journal of Machine Learning Research*, **15**, 3065–3105.

[12] Meinshausen, Nicolai, & Bühlmann, Peter. 2006. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, **34**(3), 1436–1462.

[13] Silander, Tomi, & Myllymaki, Petri. 2006. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*.

[14] van de Geer, Sara, & Bühlmann, Peter. 2013. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, **41**(2), 536–567.