Learning Linear Bayesian Networks with Latent Variables

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Department of EECS, University of California, Irvine

Daniel Hsu DAHSU@MICROSOFT.COM
Microsoft Research New England

Adel Javanmard ADELJ@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Sham M. Kakade SKAKADE@MICROSOFT.COM
Microsoft Research New England

Abstract

This work considers the problem of learning linear Bayesian networks when some of the variables are unobserved. Identifiability and efficient recovery from low-order observable moments are established under a novel graphical constraint. The constraint concerns the expansion properties of the underlying directed acyclic graph (DAG) between observed and unobserved variables in the network, and it is satisfied by many natural families of DAGs that include multi-level DAGs, DAGs with effective depth one, as well as certain families of polytrees.

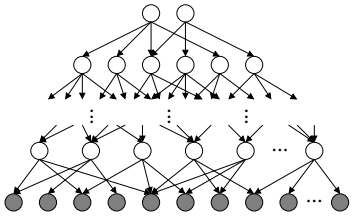
1. Introduction

It is widely recognized that incorporating latent or hidden variables is a crucial aspect of modeling. Latent variables can provide a succinct representation of the observed data through dimensionality reduction; the possibly many observed variables are summarized by fewer hidden effects. Further, they are central to predicting causal relationships and interpreting the hidden effects as unobservable concepts. For instance in sociology, human behavior is affected by abstract notions such as social attitudes, beliefs, goals and plans. As another example, medical knowledge is organized into casual hierarchies of invading organisms, physical disorders, pathological states and symptoms, and only the symptoms are observed.

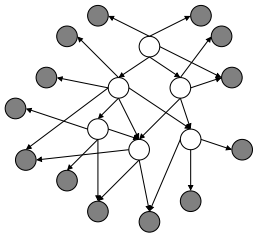
Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

In addition to incorporating latent variables, it is also important to model the complex dependencies among the variables. A popular class of models for incorporating such dependencies are the Bayesian networks, also known as belief networks. They incorporate a set of causal and conditional independence relationships through directed acyclic graphs (DAG) (Pearl, 1988). These models are widely applicable to a number of fields such as artificial intelligence, computational biology, and economics, to name a few.

An important statistical task is to learn such latent Bayesian networks from observed data. This involves discovery of the hidden variables, structure estimation (of the DAG) and estimation of the model parameters. Typically, in the presence of hidden variables, the learning task suffers from identifiability issues since there may be many models which can explain the observed data. In order to overcome indeterminacy issues, one must restrict the set of possible models. We establish novel criteria for identifiability of latent DAG models using only low order observed moments (second/third moments). We introduce a graphical constraint which we refer to as the *expansion property*. Roughly speaking, expansion property states that every subset of hidden nodes has "enough" number of outgoing edges, so they have a noticeable influence on the observed nodes, and thus on the samples drawn from the joint distribution of the observed nodes. This notion implies new identifiability and learning results for DAG structures. More specifically, we show that under this constraint, some broad families of DAG models with hidden variables, including multi-level DAGs and DAGs with effective depth



(a) Multi-level DAG



(b) DAG with effective depth one

Figure 1: A multi-level DAG and a DAG with effective depth one (observed nodes are shaded).

one, which includes (a subset of) trees and polytrees[1] satisfy this constraint and are thus, identifiable from only second and third observed moments. In addition, we propose novel and efficient algorithms for the learning task which leverage on the ideas from sparse recovery and dictionary learning (Spielman et al., 2012) as well as from spectral methods for inverse moment problems (Anandkumar et al., 2012a).

2. Model and outline of the results

2.1. Notation

We write $\|v\|_p$ for the standard ℓ^p norm of a vector v. Specifically, $\|v\|_0$ denotes the number of non-zero entries in v. Also, $\|M\|_p$ refers to the induced operator norm on a matrix M. For a matrix M and set of indices I, J, we let M_I denote the submatrix containing just the rows in I and $M_{I,J}$ denote the submatrix formed by the rows in I and columns in J. For a vector v, $\mathrm{supp}(v)$ represents the positions of non-zero entries of v. We use e_i to refer to the i-th standard basis element, *e.g.*, $e_1 = (1, 0, \ldots, 0)$. For a matrix M we let $\mathrm{Row}(M)$ (similarly $\mathrm{Col}(M)$) denote the span of its rows (columns). For a set S, $|S|$ is its cardinality. We use the notation $[n]$ to denote the set $\{1, \ldots, n\}$. For a vector v, $\mathrm{diag}(v)$ is a diagonal matrix with the elements of v on the diagonal. For a matrix M, $\mathrm{diag}(M)$ is a diagonal matrix with the same diagonal as M.

[1] A polytree is a directed acyclic graph where ignoring the directions, the graph is a tree.

2.2. Model

We define a *DAG model* as a pair $(\mathcal{G}, \mathbb{P}_\theta)$, where $\mathbb{P}_\theta$ is a joint probability distribution, parameterized by θ, on n variables $x := (x_1, \ldots, x_n)$ that is Markov with respect to a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{1, \ldots, n\}$ (Lauritzen, 1996). More specifically, the joint probability $\mathbb{P}_\theta(x)$ factors as

$$\mathbb{P}_\theta(x) = \prod_{i=1}^{n} \mathbb{P}_\theta(x_i | x_{\mathsf{PA}_i}), \tag{1}$$

where $\mathsf{PA}_i := \{j \in \mathcal{V} : (j, i) \in \mathcal{E}\}$ denotes the set of parents of node i in $\mathcal{G}$.

The learning task involving DAG models can be described as: *Given i.i.d. samples generated from the joint distribution $\mathbb{P}_\theta$ over x_S for some $S \subseteq \mathcal{V}$, recover (some part of) the graph structure $\mathcal{G}$ and estimate the model parameter θ.*

We consider DAG $\mathcal{G} = (\mathcal{V}_{\mathrm{obs}} \cup \mathcal{V}_{\mathrm{hid}}, \mathcal{E})$ with observed nodes $\mathcal{V}_{\mathrm{obs}} = \{x_1, \ldots, x_n\}$ and hidden nodes $\mathcal{V}_{\mathrm{hid}} = \{h_1, \ldots, h_k\}$. Let ε_i be the noise variable associated with x_i, for $i \in [n]$ and denote the variance of ε_i by $\sigma^2_{\varepsilon_i} > 0$. Throughout we use the notation $h := (h_1, \ldots, h_k)$, $x := (x_1, \ldots, x_n)$ and $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)$. The noise terms ε are assumed to be uncorrelated. The class of models considered are specified by the following assumptions.

Condition 1 (Linear model)**.** *The observed and hidden variables obey the model*[2]

$$x_i = \sum_{j \in \mathsf{PA}_i} a_{ij} h_j + \varepsilon_i, \quad \text{for } i \in [n], \tag{2}$$

where $\{\varepsilon_i\}$ are uncorrelated and are independent from $\{h_j\}$. Furthermore, the hidden variables are linearly independent, i.e.*, with probability one, if $\sum_{i \in [k]} \alpha_i h_i = 0$, then $\alpha_i = 0$, for all $i \in [k]$.*

We note that without a non-degeneracy assumption on the hidden variables there is no hope of distinguishing different hidden nodes.

Notice that the structure of $\mathcal{G}$ is defined by the non-zero coefficients in Eq. (2). Therefore, there is no edge among the observed nodes. We define $A \in \mathbb{R}^{n \times k}$ by letting the (i, j) entry be a_{ij} if $j \in \mathsf{PA}_i$ and zero otherwise. We refer to matrix A as the *coefficient matrix.*

Remark 2.1. *The linear relationships described above can be thought of as linear structural equation models (SEM). In general, an SEM is defined by a collection of equations*

$$z_i = f_i(z_{\mathsf{PA}_i}, \varepsilon_i), \tag{3}$$

[2] Without loss of generality, assume that x_i, ε_i, h_j are all zero mean.

with z_i be the variables associated to the nodes. Recently, there has been some progress on the identifiability problem of SEMs in the fully observed setting (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2011; Peters & Bühlmann, 2012). This paper can be viewed as a contribution to the problem of identifiability and learning SEMs with latent variables.

We now describe sufficient conditions under which the linear DAG model with hidden variables becomes identifiable. Given observations x, note that we can only hope to identify the columns of matrix A up to permutation because the model is unchanged if one permutes the hidden variables h and the columns of A correspondingly. Moreover, the scale of each column of A is also not identifiable. To see this, observe that Eq. (2) is unaltered if we both rescale all the coefficients $\{a_{ij}\}_{j\in[k]}$ and appropriately rescale the variable h_i. Without further assumptions, we can only hope to recover a certain canonical form of A, defined as follows:

Definition 2.2. *We say A is in a* canonical form *if for each $j \in [k]$, $\sigma^2_{h_j} = \mathbb{E}[h_j^2] = 1$. In particular, the transformation $A \leftarrow A\,\mathrm{diag}(\sigma_{h_1}, \sigma_{h_2}, \ldots, \sigma_{h_k})$ and the corresponding rescaling of h place A in canonical form and the distribution over x_i, $i \in [n]$, is unchanged.*

Furthermore, observe that the canonical A is only specified up to sign of each column since any sign change of column i does not alter the variance of h_i.

We now discuss a rank condition on the coefficient matrix A.

Condition 2 (Rank condition)**.** *There exists a fixed partition $\mathcal{P}$ of $[n]$ such that $|\mathcal{P}| = 3$, and A_I has full column rank for all $I \in \mathcal{P}$.*

Since $\mathrm{rank}(A_I) = k$, for $I \in \mathcal{P}$, we have as a consequence $n \geq |\mathcal{P}|\, k = 3k$. Therefore, it essentially states that the number of hidden nodes should be at most one third of the observed ones. In most applications, we are looking for *a few* number of hidden effects that can represent the statistical dependence relationships among the observed nodes. Thus the rank condition is reasonable in these cases. As we will see later, due to this assumption we can extract the noise term from the observed moments.

We proceed by defining the *expansion property* of a graph which plays a key role in establishing our identifiability results.

Definition 2.3. *Let $\mathcal{H}(\mathcal{V}_1, \mathcal{V}_2)$ be a bipartite DAG with parts $\mathcal{V}_1$ and $\mathcal{V}_2$, and edges directed from $\mathcal{V}_1$ to $\mathcal{V}_2$. We say that $\mathcal{H}(\mathcal{V}_1, \mathcal{V}_2)$ satisfies the* expansion property *if for any subset $S \subseteq \mathcal{V}_1$, with $|S| \geq 2$, we have $|\mathrm{N}(S)| \geq |S| + d_{\max}$, where $\mathrm{N}(S) := \{i \in \mathcal{V}_2 : (j, i) \in \mathcal{E}$ for some $j \in S\}$ is the set of the neighbors of S and $d_{\max}$ is the maximum degree of nodes in $\mathcal{V}_1$.*

Condition 3 (Graph expansion)**.** *Let $\mathcal{H}(\mathcal{V}_{\mathrm{hid}}, \mathcal{V}_{\mathrm{obs}})$ denote the graph formed by the edges between $\mathcal{V}_{\mathrm{hid}}$ and $\mathcal{V}_{\mathrm{obs}}$. Then, $\mathcal{H}(\mathcal{V}_{\mathrm{hid}}, \mathcal{V}_{\mathrm{obs}})$ has the expansion property.*

The last condition is a generic assumption on the entries of matrix A. We first define the *parameter genericity property* for a matrix.

Definition 2.4. *We say that matrix $M \in \mathbb{R}^{n\times k}$ has the* parameter genericity property *if for any $v \in \mathbb{R}^k$ with $\|v\|_0 \geq 2$, the following holds true.*

$$\|Mv\|_0 > |\mathrm{N}_M(\mathrm{supp}(v))| - |\,\mathrm{supp}(v)|, \tag{4}$$

where for a set $S \subseteq [k]$, $\mathrm{N}_M(S) := \{i \in [n] : M_{ij} \neq 0$ for some $j \in S\}$.

Condition 4 (Parameter genericity)**.** *The coefficient matrix A has the parameter genericity property.*

This is a mild generic condition. More specifically if the entries of an arbitrary fixed matrix M are perturbed independently, then it satisfies the above generic property with probability one.

Remark 2.5. *Fix any matrix $M \in \mathbb{R}^{n\times k}$. Let $Z \in \mathbb{R}^{n\times k}$ be a random matrix such that $\{Z_{ij} : M_{ij} \neq 0\}$ are independent random variables, and $Z_{ij} \equiv 0$ whenever $M_{ij} = 0$. Assume each variable is drawn from a distribution with uncountable support. Then*

$$\mathbb{P}(M + Z \text{ does not satisfy Condition 4}) = 0. \tag{5}$$

The proof of Remark 2.5 is available in the long version of this paper (Anandkumar et al., 2012b).

2.3. Summary of contributions

We establish identifiability of different classes of linear DAG models from the observed data, and also propose efficient algorithms for the learning task. In the following, we summarize our identifiability results and the proposed algorithms.

Identifiability. Our core result is the following.

Core result. Under the model assumptions in Section 2.2, one can identify the coefficient matrix A from the second order moment $\mathbb{E}[xx^\top]$, without additional assumptions on the dependency relationships among the hidden nodes.

This result shows how the graph expansion property enables the identifiability of connectivity structure between the set of hidden nodes and the set of observed

nodes for a general DAG. It is worth noting that the result is obtained using only the second order moments. If the hidden nodes obey a Gaussian joint distribution, then so do the observed nodes and the second moment completely characterizes their joint distribution. But in general, the second moment provides strictly smaller amount of information than the entire joint distribution. This makes our result robust to the noises in the observations as it relies on them only through the second moment.

We next consider two ensembles of DAG models, namely multi-level DAGs and DAGs with effective depth one. Building upon our core result, we show that for these ensembles the induced model among the hidden nodes is also identifiable.

Multi-level DAGs. This ensemble contains graphs with a hierarchal structure. The nodes of a multi-level DAG can be partitioned into levels $L_1, \ldots, L_m$, such that there is no edge within a level and all the edges are between nodes in level L_i and the nodes in the adjacent levels L_{i-1} and L_{i+1} (see Fig. 1(a) for an illustration). Assuming that the induced model between levels L_i and L_{i+1} obeys the conditions in Section 2.2 for $i \in [m-1]$, we show that the entire model can be learned in a sequential manner.

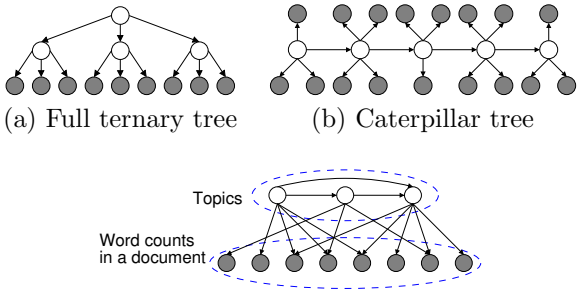
DAGs with effective depth one. A DAG has effective depth one if any hidden node has at least one observed neighbor (See Fig. 1(b) for an illustration). Now suppose that the dependence relationships among the hidden nodes are also linear and are described as follows:

$$h_j = \sum_{\ell \in \mathsf{PA}_j} \lambda_{j\ell} h_\ell + \eta_j, \quad \text{for } j \in [k], \tag{6}$$

where $\{\eta_j\}_{j\in[k]}$ denote the noise terms. For models in this class, we use Excess Correlation Analysis (ECA) (Anandkumar et al., 2012a) to learn the model from the third order moment of the observed variables. Here, we assume that the noise variables at the hidden nodes are non-Gaussian (*e.g.*, they have non-zero third moment or excess kurtosis).

Our presentation focuses on using exact (population) observed moments to emphasize the correctness of the methodology. However, "plug-in" moment estimates can be used with sampled data.

Learning algorithm. The above results already imply identifiability of the aforementioned DAG models through exhaustive search. We also present some conditions on the coefficient matrix A, under which we can efficiently learn the columns of A from the second order moment, by solving a set of convex optimization problems. This leads to efficient algorithms for learning multi-level DAGs and DAGs with effective depth one (Algorithm 1 and Algorithm 2).



(a) Full ternary tree (b) Caterpillar tree

(c) Graph representation for topic model

Figure 2: Examples of graphs from the ensembles of multi-level DAGs and DAGs with effective depth one.

Examples. It is useful to consider some concrete examples of multi-level DAGs and DAGs with effective depth one, which satisfy the expansion property. Using the results of this paper, under the rank condition and the parameter genericity property for matrix A, these models are identifiable.

Full d-regular trees. These are tree structures in which every node other than the leaves has d children. These are included in the ensemble of multi-level DAGs and it is immediate to see that for $d \geq 3$, the model can be identified under the described model in Section 2.2. (Note that $d \geq 2$ suffices for expansion property but $d \geq 3$ is necessary for the rank condition.) See Fig. 2(a) for an illustration of a full ternary tree with latent variables.

Caterpillar trees. These are tree structures in which all the leaves are within distance one of a central path. See Fig. 2(b) for an illustration. These structures have effective depth one. Let $d_{\max}$ and $d_{\min}$ respectively denote the maximum and the minimum number of leaves connected to a fixed node on the central path. It is immediate to see that if $d_{\min} \geq d_{\max}/2 + 1$, the structure has the expansion property.

Random bipartite graphs. Consider bipartite graphs with hidden nodes in one part and observed nodes in the other part. Each edge (between the two parts) is included in the graph with probability θ, independent from every other edge. It is easy to see that, for any set $S \subseteq [k]$, the expected number of its neighbors is $\mathbb{E}|\mathrm{N}(S)| = n(1-(1-\theta)^{|S|})$. Also, the expected degree of the hidden nodes is θn. Now, by applying a Chernoff bound, one can show that these graphs have the expansion property with high probability, if $k \leq \theta n/2$, *i.e.*, with probability converging to one as $n \to \infty$.

Application to correlated topic models. An im-

portant application of the results of this paper is in estimating topic models with correlated topics. Topic models are a popular family of mixture models that incorporate latent variables, the topics, to explain the observed co-occurrences of words in documents. Each document has a mixture of active topics and each active topic determines the occurrence of words in the document. A topic model can be viewed as a bipartite DAG with topics in one part and the observed nodes in the other part. See Fig. 2(c) for an illustration. (As an example, one may think of the i-th observed variable as the word counts in the i-th sentence of a document.) Using this representation, estimating the topics from the document is exactly the learning problem of the corresponding DAG. Existing work on estimating topic models provide results for certain distributions over the topics. For instance, in independent component analysis (ICA), the topics are assumed to be independent, while in Latent Dirichlet Allocation (LDA), a Dirichlet prior is assigned to the distribution of topics in documents. However, it has been observed empirically that correlated topic models provide better fit for document modeling (Blei & Lafferty, 2007; Li & McCallum, 2006). A popular correlated topic model, termed as *Pachinko allocation* involves multi-level DAGs for modeling word dependencies. We can now efficiently learn a rich class of similar correlated topic models.

2.4. Our techniques

Our proof techniques rely on ideas and tools developed in dictionary learning, matrix decomposition, and method of moments. We briefly explain our techniques and their relations to these areas.

Matrix decomposition into diagonal and low-rank parts. To prove our core result, we first observe that under the linear model, $\mathbb{E}[xx^\top]$ is the sum of a low-rank matrix and a diagonal one:

$$\mathbb{E}[xx^\top] = A\mathbb{E}[hh^\top]A^\top + \mathbb{E}[\varepsilon\varepsilon^\top].$$

We prove that under the rank condition (Condition 2), $\mathbb{E}[xx^\top]$ can be decomposed into its low-rank component $A\mathbb{E}[hh^\top]A^\top$ and its diagonal component $\mathbb{E}[\varepsilon\varepsilon^\top]$. This means that we can remove the noise contribution from the second order moment. Moreover, $\mathrm{rank}(A\mathbb{E}[hh^\top]A^\top) = k$ gives the number of hidden nodes. We propose a simple algorithm (SUBROUTINE) for this decomposition.[3]

Dictionary learning. We proceed by showing that using the graph expansion property (Condition 3), one can recover A from the low-rank part $A\mathbb{E}[hh^\top]A^\top$, obtained from the decomposition of the observed covariance matrix, as described above. To prove this claim, we leverage the ideas developed by Spielman et al. (2012) for the dictionary learning problem. Spielman et al. consider the problem of learning sparsely used dictionaries with an invertible dictionary and a random, sparse coefficient matrix, Bernoulli-Gaussian and Bernoulli-Radamacher models. They establish that the dictionary and the coefficient matrix can be learned from exact measurements. The gist of the idea is that under the above conditions, the row space of the coefficient matrix is the same as that of the measurements matrix. The rows of the coefficient matrix are then the sparsest vectors in the corresponding space.

Notice that here we are in the same situation. Since $\mathbb{E}[hh^\top]$ and A have full column rank, we have $\mathrm{Col}(A) = \mathrm{Col}(A\mathbb{E}[hh^\top]A^\top)$. However, in contrast to the dictonary learning setting of Spielman et al. (2012), the coefficient matrix A is not generated from a probabilistic model. We introduce the graph expansion property as the underlying notion which makes the recovery of A possible. In fact, it can be shown that the probabilistic models considered by Spielman et al. possess this property almost surely. Our core result (identifiability of A) is established by showing that, under the expansion property for the model, the columns of A are the sparsest vectors in $\mathrm{Col}(A\mathbb{E}[hh^\top]A^\top)$.

Method of moments.

For DAGs with effective depth one, observe that the hidden variables are related to each other and to the noise terms $\{\eta_j\}_{j\in[k]}$ via linear equations (6). Define $\Lambda \in \mathbb{R}^{k\times k}$ by letting the (i,j) entry be λ_{ij} if $j \in \mathsf{PA}_i$ and zero otherwise. Solving for the hidden variables h_j, we have $h = (I-\Lambda)^{-1}\eta$, with $\eta := (\eta_1, \ldots, \eta_k)$. The observed variables are also related to the hidden ones via the coefficient matrix A. The idea is to consider an equivalent DAG model obtained by suppressing the hidden nodes h_j and treating the noise terms η_j as the new uncorrelated hidden variables. The observed variables x_i are then related to the new hidden variables through the matrix $A(I-\Lambda)^{-1}$. Next, we apply ECA method of Anandkumar et al. (2012a) to learn $A(I-\Lambda)^{-1}$ from the second and third order moments of the observed variables. ECA is based on two singular value decompositions: the first SVD whitens the data (using second moment) and the sec-

[3] It should be noted that additive matrix decompositions into low-rank and diagonal (or sparse) terms have been considered in previous work (Chandrasekaran et al., 2011; Hsu et al., 2011; Saunderson et al., 2012). Using the techniques of Saunderson et al. (2012), we can relax Condition 2 to $k \le n/2$, but only by imposing additional strong incoherence conditions on the low-rank component.

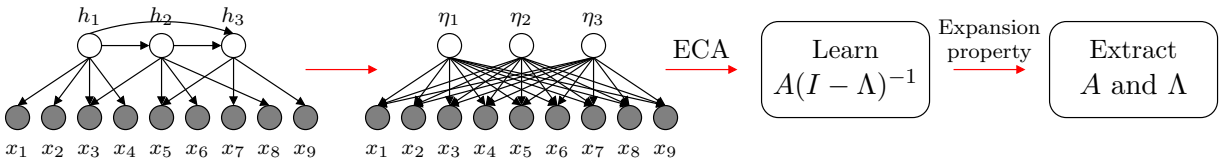


Figure 3: The high-level idea of the technique used for learning DAGs with effective depth one. In the leftmost graph (original DAG) the hidden nodes depend on each other through the matrix Λ and the observed variables depend on the hidden nodes through the coefficient matrix A. We consider an equivalent DAG with new uncorrelated hidden variables η_j (these are in fact the noise terms at the hidden nodes in the previous model). Here, the observed variables depend on the hidden ones through the matrix $A(I-\Lambda)^{-1}$. Applying ECA method, we learn this matrix from the (second and third order) observed moments. Finally, using the expansion property of the connectivity structure between the hidden part and the observed part, we extract A and Λ from $A(I-\Lambda)^{-1}$.

ond SVD uses the third moment to find directions which exhibit information that is not captured by the second moment. Finally, in order to identify the dependence structure among the hidden nodes (matrix Λ), we use the expansion property to extract A and Λ from $A(I-\Lambda)^{-\top}$. The high-level idea is depicted in Fig. 3.

2.5. Related work

The problem of identifiability and learning graphical models from distributions has been the object of intensive investigation in the past years and has been studied in different research communities. This problem has proved important in a vast number of applications, such as computational biology, economics, sociology, and computer vision (see, *e.g.*, Durbin et al., 1998; Zellner, 1971; Bollen, 1989; Choi et al., 2010). The learning task has two main ingredients: structure learning and parameter estimation.

Structure estimation has been extensively studied in the recent years. It is well known that maximum likelihood estimation in fully observed tree models is tractable (Chow & Liu, 1968). However, for general models, structure learning is NP-hard even when there are no hidden variables. The main approaches for structure estimation are score-based methods, local tests, and convex relaxation methods. Score-based methods (*e.g.*, Chickering, 2003) find the graph structure by optimizing a score, like Bayesian Independence criterion (BIC), in a greedy manner. Local test approaches attempt to build the graph based on local statistical tests on the samples, both for directed and undirected graphical models (*e.g.*, Spirtes et al., 2000; Bresler et al., 2008). Convex relaxation approaches have also been considered for structure estimation (*e.g.*, Ravikumar et al., 2010).

In the presence of latent variables, structure learning becomes more challenging. A popular class of latent variable models are latent trees, for which efficient algorithms have been developed (*e.g.*, Erdös et al., 1999; Anandkumar et al., 2011). Recently, approaches have been proposed for learning (undirected) latent graphical models with long cycles in certain parameter regimes (Anandkumar & Valluvan, 2012). Chandrasekaran et al. (2012) estimate latent Gaussian graphical models using convex relaxation approaches via sparse + low rank matrix decomposition. Silva et al. (2006) study linear latent DAG models and propose methods to (1) find clusters of observed nodes that are separated by a single latent common cause; and (2) find features of the Markov Equivalence class of causal models for the latent variables. Their model allows for undirected edges between the observed nodes. Ali et al. (2005) characterizes equivalence classes of DAG models when there are latent variables. However, the focus is on constructing an equivalence class of DAG models, given a member of the class. In contrast, we focus on developing efficient learning methods for latent DAGs.

For parameter estimation with hidden variable models, the traditional approach is expectation maximization (EM) algorithm, which finds a local maximizer of the likelihood. Unfortunately, optimality and recovery guarantees are generally lacking for EM, even when the model is correct. Another approach is to constrain the dependency structure among the hidden nodes. For instance, in *independent component analysis* (ICA) (Hyvärinen et al., 2001), it is assumed that the latent variables obey a product distribution and hence in the corresponding graph model there is no edge between the latent variables (there are only directed edges from latent nodes to the observed nodes). Several generalizations of ICA have also been developed that allow some dependent components (*e.g.*, Bach & Jordan, 2003; Theis, 2007). Anandkumar et al.

(2012a) considers latent variables to be drawn from a Dirichlet distribution, relevant in topic modeling (Blei et al., 2003), and obtains parameter estimates via the method of moments. In this work, we also use the method of moments to establish identifiability and efficient recovery for DAG models.

3. Main results

In this section, we state our identifiability results and algorithms for learning the DAG models with latent variables. Due to space limitations, we omit proofs and most technical details, and refer the interested reader to (Anandkumar et al., 2012b).

3.1. Learning the coefficient matrix A

Our core identifiability result is the following theorem.

Theorem 3.1. *Let $\Sigma := \mathbb{E}[xx^\top]$ be the second order moment of the observed variables. For the model described in Section 2.2 (Conditions 1, 2, 3, 4), all columns of A are identifiable from Σ.*

As shown in the proof, columns of A are in fact the sparsest vectors in the space $\mathrm{Col}(A\mathbb{E}[hh^\top]A^\top)$. This result already implies identifiability of A via an exhaustive search, which is an interesting result in its own right. The following theorem provides some conditions under which the columns of A can be identified by solving a set of convex optimization problems. Before stating the theorem, we need to establish some notations.

For $i \in [n]$, we define $\mathrm{N}_i := \{j \in [k] : A_{ij} \neq 0\}$ and $\mathrm{N}_i^2 := \{l \in [n] : A_{lj} \neq 0 \text{ for some } j \in \mathrm{N}_i\}$. Similarly, for $j \in [k]$, define $\mathrm{N}_j := \{i \in [n] : A_{ij} \neq 0\}$ and $\mathrm{N}_j^2 := \{l \in [k] : A_{il} \neq 0 \text{ for some } i \in \mathrm{N}_j\}$. We use superscript c to denote the set complement.

Theorem 3.2. *Suppose that in each row of A, there is a gap between the maximum and the second maximum absolute values. For $i \in [n]$, let π_i be a permutation such that $|a_{i,\pi_i(1)}| \geq |a_{i,\pi_i(2)}| \geq \cdots \geq |a_{i,\pi_i(k)}|$, and $|a_{i,\pi_i(2)}|/|a_{i,\pi_i(1)}| \leq 1 - \gamma_i$, for some $\gamma_i > 0$. Further suppose that $[k] \subseteq \{\pi_1(1), \ldots, \pi_n(1)\}$. In words, each column contains at least one entry that has the maximum absolute value in its row. If the following conditions hold true for $i \in [n]$, then* Algorithm 1 *returns the rows of A in canonical form.*

(i) $\|A_{(\mathrm{N}_i^2)^c,(\mathrm{N}_i)^c}\, v\|_1 > \|A_{\mathrm{N}_i^2,(\mathrm{N}_i)^c}\, v\|_1$ *for all non-zero vectors* $v \in \mathbb{R}^{|(\mathrm{N}_i)^c|}$.

(ii) $\|A_{(\mathrm{N}_j)^c,\mathrm{N}_i\setminus j}\, v\|_1 > \|A_{\mathrm{N}_j,\mathrm{N}_i\setminus j}\, v\|_1 + (1-\gamma)\|A_{\mathrm{N}_j,j}\|_1 \|v\|_1$ *for all* $j \in \mathrm{N}_i$ *and all non-zero vectors* $v \in \mathbb{R}^{|\mathrm{N}_i|-1}$.

Subroutine: Decomposition of a matrix into its low-rank and diagonal parts.

Input: Matrix $C = AB^\top + D$, with $A, B \in \mathbb{R}^{n\times k}$, $D \in \mathbb{R}^{n\times n}$ diagonal, and partition $\mathcal{P}$ of $[n]$.
Output: Diagonal D and low-rank $L = AB^\top$ parts.

for each $I \in \mathcal{P}$ **do**
  Choose distinct $J, K \in \mathcal{P}\setminus\{I\}$.
  Let $U_I \in \mathbb{R}^{|I|\times k}$ be the matrix of left singular vectors of $C_{I,J}$.
  Let $V_J \in \mathbb{R}^{|J|\times k}$ be the matrix of right singular vectors of $C_{I,J}$.
  Let $U_K \in \mathbb{R}^{|K|\times k}$ be the matrix of left singular vectors of $C_{K,J}$.
  Set $A_I B_I^\top = C_{I,J} V_J (U_K^\top C_{K,J} V_J)^{-1} U_K^\top C_{K,I}$.
  Set $D_{I,I} = C_{I,I} - A_I B_I^\top$.
return D and $L = C - D$.

Algorithm 1: Recovering columns of coefficient matrix A from the second order moment Σ.

Input: Second order moment of observed variables Σ.
Output: Columns of A up to permutation.

Find a partition $\mathcal{P}$ of $[n]$ such that $|\mathcal{P}| = 3$ and $\mathrm{rank}(\Sigma_{I,J}) = k$ for distinct $I, J \in \mathcal{P}$.
Let L be the low-rank part returned by Subroutine$(\Sigma, \mathcal{P})$.
for each $i \in [n]$ **do**
  Solve the optimization problem

$$\min_w \ \|L^{1/2}w\|_1 \qquad \text{subject to } (e_i^\top L^{1/2})w = 1.$$

  Set $s_i = L^{1/2}w$, and let $\mathcal{S} = \{s_1, \ldots, s_n\}$.
for each $j = 1, \ldots, k$ **do**
  repeat
    Let v_j be an arbitrary element in $\mathcal{S}$.
    Set $\mathcal{S} = \mathcal{S}\setminus\{v_j\}$.
  until $\mathrm{rank}([v_1|\cdots|v_j]) = j$
Set $\tilde{A} = [v_1|\cdots|v_k]$.
Let $\tilde{B}$ be a left inverse for $\tilde{A}$, *i.e.*, $\tilde{B}\tilde{A} = I_{k\times k}$.
return Columns of $\tilde{A}(\mathrm{diag}(\tilde{B}L\tilde{B}^\top))^{1/2}$.

Algorithm 1 is essentially the ER-SpUD algorithm of Spielman et al. (2012) for exact recovery of sparsely-used dictionaries, but the technical result and application in Theorem 3.2 are novel.

According to Theorem 3.1, we can learn the coefficient matrix A of the model without any assumption on the dependence relationships among the hidden nodes. (We only need the non-degeneracy assumption in Condition 1, which requires the hidden variables to be linearly independent with probability one.)

Note that the coefficient matrix A does not completely

specify the distribution, as the h_i's are not necessarily statistically independent, and we can hope to learn the correlation structure among the h_i's. We next consider two families of DAG models, namely multi-level DAGs and DAGs with effective depth one. For these families, we proceed further and prove identifiability of the entire model.

3.2. Multi-level DAGs

We first formally define multi-level DAGs.

Definition 3.3. *A multi-level DAG model is a model with the following graph structure. The nodes of the graph can be partitioned into levels $L_1, \ldots, L_m$ such that there is no edge between the nodes within one level and all the edges are between nodes in adjacent levels, (L_i, L_{i+1}) for $i \in [m-1]$. Furthermore, the edges are directed from L_i to L_{i+1}. The nodes in level L_m correspond to the observed nodes and other levels contain the hidden nodes.*

The next theorem concerns identifiability of linear multi-level DAGs. More specifically, consider a multi-level DAG model and let $\mathcal{G}_i$ be the induced graph with nodes $L_i \cup L_{i+1}$ and suppose that the induced model between levels L_i and L_{i+1} satisfies the model conditions described in Section 2.2 with coefficient matrix A_i, for $i \in [m-1]$: A_i has the rank condition (Condition 2) and parameter genericity property (Condition 4), and (bipartite) graph $\mathcal{G}_i$ has the expansion property (Condition 3).

Theorem 3.4. *Consider a multi-level DAG with levels $L_1, \ldots, L_m$ and suppose that the induced model between levels L_i and L_{i+1} satisfies the model conditions described in Section 2.2 with coefficient matrix A_i, for $i \in [m-1]$. Then all columns of A_i are identifiable for $i \in [m-1]$ from the second order moment of the observed variables Σ. Therefore, the entire DAG is identifiable up to permuting the nodes within each level.*

Remark 3.5. *By the definition of a multi-level DAG, the hidden nodes in level L_1 are independent. Now consider the case that the nodes in L_1 have arbitrary dependence relationships. By using the same argument as in the proof of Theorem 3.4, we can still learn all the coefficient matrices A_i and the second order moment of the nodes in L_1.*

3.3. DAGs with effective depth one

Another important subclass of DAGs are those with effective depth one.

Definition 3.6. *The* effective depth *of a DAG model with hidden nodes is the maximum graph distance between a hidden node and its closest observed node.*

In particular, in a DAG with effective depth one every hidden node has at least one observed neighbor.

Recall that the observed and the hidden nodes obey the linear model in Eq. (2), which in vector form reads $x = Ah + \varepsilon$. Let $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{k \times k}$ be the matrix with $\lambda_{ij} = 0$ if $j \notin \mathsf{PA}_i$. We assume further that the hidden variables obey the linear model in Eq. (6), *i.e.*, $h = \Lambda h + \eta$.

As described in Section 2.2, without loss of generality, we assume that hidden variables h_j, the observed variables x_i and the noise terms ε_i, η_j are all zero mean. We also denote the variances of ε_i and η_j by $\sigma^2_{\varepsilon_i}$ and $\sigma^2_{\eta_j}$, respectively. Let μ_{ε_i} and μ_{η_j} respectively denote the third moment of ε_i and η_j, *i.e.*, $\mu_{\varepsilon_i} := \mathbb{E}[\varepsilon_i^3]$ and $\mu_{\eta_j} := \mathbb{E}[\eta_j^3]$. Define the skewness of η_j as $\gamma_{\eta_j} := \frac{\mu_{\eta_j}}{\sigma^3_{\eta_j}}$. Finally, denote the second and third order correlations of the observed variables by $\Sigma := \mathbb{E}[xx^\top]$ and $\Psi := \mathbb{E}[x \otimes x \otimes x]$, where $\otimes$ denotes the tensor product.

Theorem 3.7. *Consider a DAG model with effective depth one, which satisfies the model conditions described in Section 2.2 and the hidden variables are related through linear equations (6). If the noise variables η_j have non-zero skewness for $j \in [k]$, then the DAG model is identifiable from Σ and Ψ.*

Furthermore, under the assumptions of Theorem 3.2, there is an efficient algorithm that returns matrices A and Λ up to a permutation of hidden nodes. The algorithm basically combines the Excess Correlation Analysis method of (Anandkumar et al., 2012a) and Algorithm 1. The details of the algorithm are given in the full version of the paper under the name of Algorithm 2.

3.4. Remark on finding the partition $\mathcal{P}$

In the full version of the paper, we show that under a weak *incoherence* assumption about A, a random partitioning of its rows into three groups satisfies Condition 2, with fixed positive probability.

Acknowledgments

The authors thank anonymous reviewers for their useful comments. A. Anandkumar acknowledges the support of NSF Award CCF 1219234, AFOSR Award FA9550-10-1-0310, and ARO Award W911NF-12-1-0404. Part of this work was completed while A. Anandkumar and A. Javanmard were visiting MSR New England.

References

Ali, R. A., Richardson, T., Spirtes, P., and Zhang, J. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In *UAI*, 2005.

Anandkumar, A. and Valluvan, R. Learning loopy graphical models with latent variables: Efficient methods and guarantees. arXiv:1203.3887, 2012.

Anandkumar, A., Chaudhuri, K., Hsu, D., Kakade, S. M., Song, L., and Zhang, T. Spectral methods for learning multivariate latent tree structure. In *NIPS*, 2011.

Anandkumar, A., Foster, D. P., Hsu, D., Kakade, S. M., and Liu, Y.-K. A spectral algorithm for latent Dirichlet allocation. In *NIPS*, 2012a.

Anandkumar, A., Hsu, D., Javanmard, A., and Kakade, S. M. Learning linear bayesian networks with latent variables. arXiv:1209.5350, 2012b.

Bach, F. R. and Jordan, M. I. Beyond independent components: trees and clusters. *JMLR*, 4:1205–1233, 2003.

Blei, D. M. and Lafferty, J. D. A correlated topic model of science. *Annals of Applied Statistics*, pp. 17–35, 2007.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Bollen, K. A. *Structural Equations with Latent Variables*. Wiley, New York, 1989.

Bresler, G., Mossel, E., and Sly, A. Reconstruction of Markov random fields from samples: some observations and algorithms. In *APPROX*, 2008.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Latent variable graphical model selection via convex optimization. *Annals of Statistics (to appear)*, 2012.

Chickering, D. M. Optimal structure identification with greedy search. *JMLR*, 3:507–554, 2003.

Choi, M. J., Lim, J. J., Torralba, A., and Willsky, A. S. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Tran. on Information Theory*, 14(3):462–467, 1968.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

Erdös, P. L., Székely, L. A., Steel, M. A., and Warnow, T. J. A few logs suffice to build (almost) all trees: Part I. *Random Structures and Algorithms*, 14:153–184, 1999.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *NIPS*, 2009.

Hsu, D., Kakade, S. M., and Zhang, T. Robust matrix decomposition with sparse corruptions. *IEEE Trans. on Inf. Theory*, 57(11):7221–7234, 2011.

Hyvärinen, A., Karhunen, J., and Oja., E. *Independent Component Analysis*. Wiley Interscience, 2001.

Lauritzen, S. *Graphical Models*. Oxford University Press, 1996.

Li, W. and McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, pp. 577–584, 2006.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems—Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Peters, J. and Bühlmann, P. Identifiability of Gaussian structural equation models with same error variances. arXiv:1205.2536v1, 2012.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In *UAI*, 2011.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

Saunderson, J., Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. arXiv:1204.1220, 2012.

Shimizu, S., Hoyer, P. O., Hyvärisen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7:2003–2030, 2006.

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. Learning the structure of linear latent variable models. *JMLR*, 7:191–246, 2006.

Spielman, D. A., Wang, H., and Wright, J. Exact recovery of sparsely-used dictionaries. arXiv:1206.5882v1, 2012.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Theis, F. J. Towards a general independent subspace analysis. In *NIPS*, 2007.

Zellner, A. *Introduction to Bayesian Inference in Econometrics*. New York: John Wiley, 2nd edition, 1971.