Learning Linear Bayesian Networks with Latent Variables

Animashree Anandkumar[1], Daniel Hsu[2], Adel Javanmard[3], and Sham M. Kakade[2]

[1]Department of EECS, University of California, Irvine
[2]Microsoft Research New England
[3]Department of Electrical Engineering, Stanford University

November 26, 2012

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.

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) [36]. They have widespread applicability in artificial intelligence [13, 19, 31, 32], in the social sciences [9, 12, 30, 37, 38, 51], and as structural equation models in economics [7, 12, 24, 38, 47, 52].

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

---

E-mail: a.anandkumar@uci.edu, dahsu@microsoft.com, adelj@stanford.edu, skakade@microsoft.com

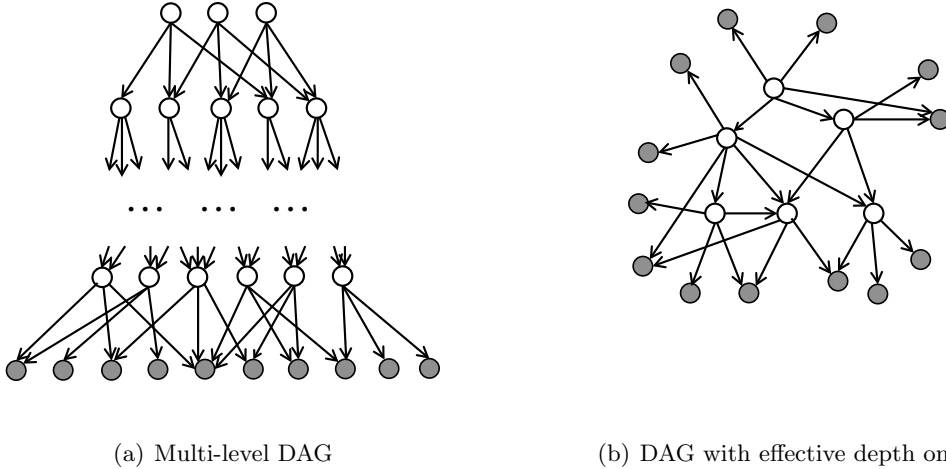(a) Multi-level DAG   (b) DAG with effective depth one

Figure 1: Illustrations of multi-level DAGs and DAGs with effective depth one. Observed nodes and hidden nodes are respectively shown by shaded and white circles. Under the expansion property for the graph and the linear dependence model (Section 2.2), we prove identifiability of these ensembles from low order moments of the observed variables.

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 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 [46] as well as from spectral methods for inverse moment problems [4].

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, $\text{supp}(v)$ represents

---

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

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.

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 $\mathcal{V} = \{1, \ldots, n\}$ [33]. 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 to x_i, for $i = 1, \ldots, n$ and denote the variance of ε_i by $\sigma_{\varepsilon_i}^2 > 0$. 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 (Linearity). *The observed and hidden variables obey the model[2]*

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

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*

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

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 [26, 39, 40, 44]. This paper can be viewed as a contribution to the problem of identifiability and learning SEMs with latent variables.

---

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

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 permute 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 $j \in [k]$, $\sigma_{h_j}^2 = \mathbb{E}[h_j^2] = 1$. In particular, the transformation $A \leftarrow A \operatorname{diag}(\sigma_{h_1}, \sigma_{h_2}, \ldots, \sigma_{h_k})$ 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 $\operatorname{rank}(A_I) = k$, 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 $\mathrm{N}(S) := \{i \in \mathcal{V}_2 : (j, i) \in \mathcal{E} \text{ for some } j \in S\}$ 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.*

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

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

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*

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

Remark 2.5 is proved in Appendix A.

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 any assumption on the dependence 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 layer and all the edges are between nodes in layer L_i and the nodes in the adjacent layers L_{i-1} and L_{i+1} (See Fig. 1(a) for an illustration). Assuming that the induced model between layers L_i and L_{i+1} obey the conditions in Section 2.2 for $i = 1, \ldots, m - 1$, 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) [4] 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.

5

(a) Full ternary tree

(b) Caterpillar tree

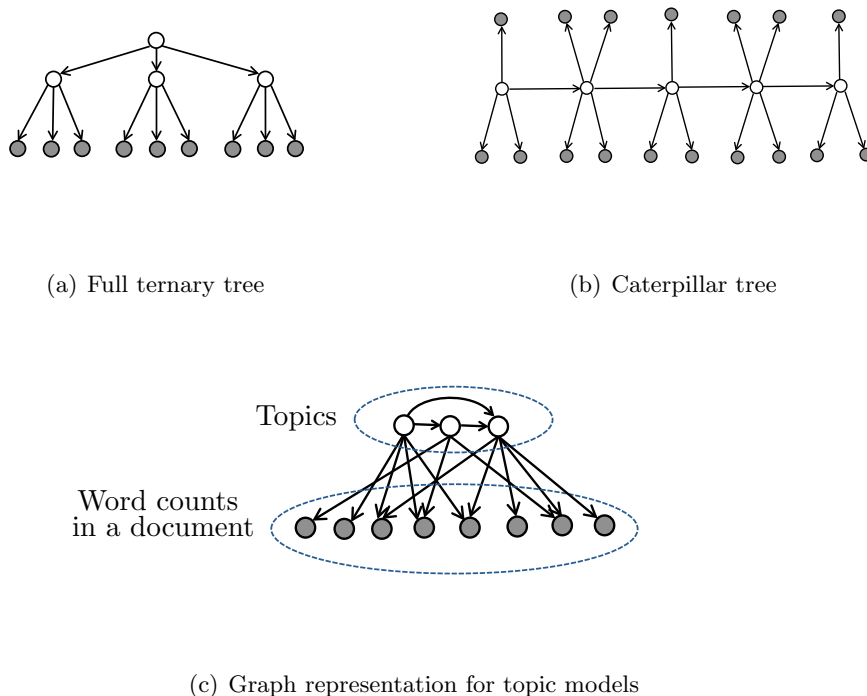(c) Graph representation for topic models

Figure 2: Concrete examples of graphs from the ensembles of multi-level DAGs and DAGs with effective depth one. Observed nodes and hidden nodes are respectively shown by shaded and white circles. Using the results of this paper, these graphs are identifiable, under the linear dependence model (Section 2.2), from second- and third-order moments of the observed variables.

To partially address the statistical efficiency of our method, note that higher-order empirical moments generally have higher variance than lower-order empirical moments, and therefore are more difficult to reliably estimate. Our techniques only involve low-order moments (up to third order). A precise analysis of sample complexity involves standard techniques for dealing with sums of i.i.d. random matrices and tensors as in [4] and is left as a future study.

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).

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 : $\mathbb{E}|N(S)| = n(1 - (1 - \theta)^{|S|})$. 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 important 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 [10, 34]. 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.

It is convenient to discuss a concrete example which further showcases the applicability of our results for topic models. Consider the linear model as described by Eqs. (2),(6) and suppose that the noise variables are independently Poisson random variables and all hidden and observed variables are Poisson. Note that sum of independent Poisson random variables is also Poisson, and therefore this is a valid model. This scenario is readily applicable for topic modeling since we can interpret each observed Poisson variable as specifying the count of a certain word, and each hidden Poisson variable as giving the count of a certain topic, and there can be arbitrary dependencies among the hidden topics. Prior to this work, even basic parameter and structural identifiability of such correlated topic models was not known. Our work gives, for the first time, a computationally efficient estimator that relies on estimation using only low-order moments.

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, $\text{rank}(A\mathbb{E}[hh^\top]A^\top) = k$ gives the number of hidden nodes. We propose a simple algorithm (Subroutine) for this decomposition.

It should be noted that additive matrix decompositions into low-rank and diagonal (or sparse) terms have been considered in previous work [16, 27, 43]. Using the techniques of [43], we can relax Condition 2 to $k \leq n/2$, but only by imposing additional strong incoherence conditions on the low-rank component.

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 in [46] for the dictionary learning problem. In [46], the authors 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 $\text{Col}(A) = \text{Col}(A\mathbb{E}[hh^\top]A^\top)$. However, in contrast to the setting in [46], 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 considered probabilistic models in [46], 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 $\text{Col}(A\mathbb{E}[hh^\top]A^\top)$.

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 $j \in \text{PA}_i$ 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 topics. The observed variables x_i are then related to the new topics through the matrix $A(I - \Lambda)^{-1}$. Next, we apply ECA method [4] to learn $A(I - \Lambda)^{-\top}$ 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 second 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
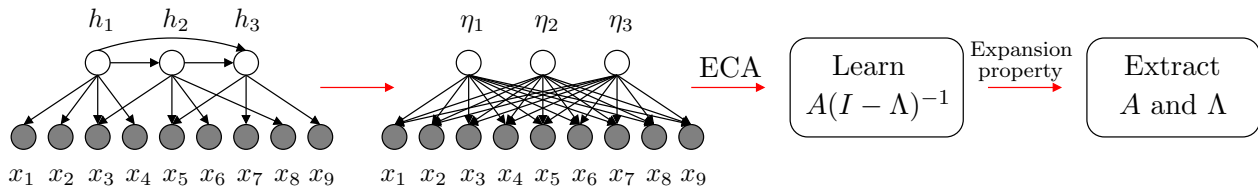
Figure 3: The high-level idea of the technique used for learning DAGs with effective depth one. In the leftmost graph (original DAG) the hidden nodes depend on each other through the matrix Λ and the observed variables depend on the hidden nodes through the coefficient matrix A. We consider an equivalent DAG with new uncorrelated topics η_j (these are in fact the noise terms at the hidden nodes). 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 $A(I - \Lambda)^{-1}$.

problem has proved important in a vast number of applications, such as computational biology [22,42], economics [7,12,24,52], sociology [9,12,30,51], and computer vision [19,32]. 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 [20]. 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 such as [17] 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 [1,5,14,25,29,48]. Convex relaxation approaches have also been considered for structure estimation, e.g. [35,41].

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 [3,18,21, 23]. Recently, approaches have been proposed for learning (undirected) latent graphical models with long cycles in certain parameter regimes [6]. In [15], latent Gaussian graphical models are estimated using convex relaxation approaches. The authors in [45] 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. In [2], equivalence class of DAG models is characterized 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) [28], 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 where clusters of dependent components are learned, *e.g.* [8] considers tree component analysis where the latent nodes form a tree model; [49] considers independent subspace analysis; and [4] considers latent variables to be drawn from latent Dirichlet allocation (LDA), relevant in topic modeling [11]. These approaches are based on the method of moments, where the observed moments are matched to those specified by the model. In this paper, we use ideas from 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.

3.1 Learning the coefficient matrix A

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.*

Theorem 3.1 is proved in Section 4.1. 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 $|a_{i,\pi_i(1)}|/|a_{i,\pi_i(2)}| \leq 1 - \gamma_i$, 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 $v \in \mathbb{R}^{|\mathrm{N}_i|-1}$.*

Theorem 3.2 is proved in Section 4.2. ALGORITHM 1 is essentially ER-SpUD presented in [46] for exact recovery of sparsely-used dictionaries; but the technical result and application 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 discussed in Condition 1 which requires that the hidden variables be linearly independent with probability one.)

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 part D and low-rank part $L = AB^\top$.

  1: **for** each $I \in \mathcal{P}$ **do**
  2:    Choose distinct $J, K \in \mathcal{P} \backslash \{I\}$.
  3:    Let $U_I \in \mathbb{R}^{|I| \times k}$ be the matrix of left singular vectors of $C_{I,J}$.
  4:    Let $V_J \in \mathbb{R}^{|J| \times k}$ be the matrix of right singular vectors of $C_{I,J}$.
  5:    Let $U_K \in \mathbb{R}^{|K| \times k}$ be the matrix of left singular vectors of $C_{K,J}$.
  6:    Set $A_I B_I^\top = C_{I,J} V_J (U_K^\top C_{K,J} V_J)^{-1} U_K^\top C_{K,I}$.
  7:    Set $D_{I,I} = C_{I,I} - A_I B_I^\top$.
  8: **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 the observed variables Σ.
Output: Columns of A up to permutation.

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

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

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

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

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 layer and all the edges are between nodes in adjacent layers, (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 layer L_m correspond to the observed nodes and other layers contain the hidden nodes.*

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 layers 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.*

Theorem 3.4 is proved in Section 4.3.

Remark 3.5. *By definition of 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

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

$$x_i = \sum_{j \in \mathsf{PA}_i} a_{ij} h_j + \varepsilon_i, \quad \text{for } i \in [n]. \tag{7}$$

Assume further that the hidden variables obey the linear model

$$h_j = \sum_{\ell \in \mathsf{PA}_j} \lambda_{j\ell} h_\ell + \eta_j, \quad \text{for } j \in [k]. \tag{8}$$

Let $\Lambda \in \mathbb{R}^{k \times k}$ be the matrix with λ_{ij} at the (i,j) entry if $j \in \mathsf{PA}_i$ and zero everywhere else.

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 $\sigma_{\varepsilon_i}^2$ and $\sigma_{\eta_j}^2$, 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:

$$\gamma_{\eta_j} := \frac{\mu_{\eta_j}}{\sigma_{\eta_j}^3}. \tag{9}$$

Finally, denote the second and third order correlations of the observed variables as:

$$\begin{aligned} \Sigma &:= \mathbb{E}[xx^\top], \\ \Psi &:= \mathbb{E}[x \otimes x \otimes x], \end{aligned} \tag{10}$$

where $\otimes$ denotes the tensor product. It is convenient to consider the projection of Ψ to a matrix as follows:

$$\Psi(\zeta) := \mathbb{E}[xx^\top \langle \zeta, x \rangle],$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner 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 (8). If the noise variables η_j have non-zero skewness for $j \in [k]$, then the DAG model is identifiable from Σ and $\Psi(\zeta)$, for an appropriate choice of ζ. Furthermore, under the assumptions of Theorem 3.2, ALGORITHM 2 returns matrices A and Λ up to a permutation of hidden nodes.*

Theorem 3.7 is proved in Section 4.4. In Theorem 3.7, we prove identifiability of DAGs with effective depth one, from the second and third order moments. A natural question is what can be done if only the second order moment is provided. The following remark states that if an oracle gives a topological ordering of the DAG structure then the model can be learned only through the second order moment and there is no need to the third order moment.

Remark 3.8. *A topological ordering of a DAG is a labeling of the nodes such that, for every directed edge (j,i), we have $j < i$. It is a well known result in graph theory that a directed graph is a DAG if and only if it admits a topological ordering. Now, consider a DAG model with effective depth one and suppose that an oracle provides us with a topological ordering of the induced DAG on the hidden nodes, i.e., for any labeling of the hidden nodes the oracle returns a permutation of the labels which is faithful to a topological ordering of the DAG. Then, the DAG model (matrices A and Λ) are identifiable from only the second order moment Σ.*

Remark 3.8 is proved in Appendix D.

Remark 3.9 (Learning fully-observed DAGs)**.** *An interesting and immediate application of the technique used in the proof of Theorem 3.7 is in learning fully-observed DAGs. Consider an arbitrary fully-observed linear DAG:*

$$x_i = \sum_{i \in \mathsf{PA}_i} \lambda_{ij} x_j + \eta_i, \quad \text{for } i \in [n], \tag{11}$$

and suppose that the noise variables η_i have non-zero skewness. Then, applying the same argument as in the proof of Theorem 3.7, we can learn the matrix $(I - \Lambda)^{-1}$ (and hence Λ) from the second and third order moments.

ALGORITHM 2: Learning DAGs with effective depth one.

Input: Vector $\theta \in \mathbb{R}^k$, observable moments Σ and Ψ as defined in Eq. (10).
Output: Columns of A, matrix Λ (in a topological ordering).
 1: **Part 1: Decomposition of Σ.**
 2:     Find a partition $\mathcal{P}$ of $[n]$ such that $|\mathcal{P}| = 3$ and $\mathrm{rank}(\Sigma_{I,J}) = k$ for distinct $I, J \in \mathcal{P}$.
 3:     Let L_Σ and D_Σ be the low-rank and the diagonal parts returned by SUBROUTINE($\Sigma, \mathcal{P}$).
 4: **Part 2: ECA.**
 5:     Find a matrix $U \in \mathbb{R}^{n \times k}$ such that $\mathrm{Col}(U) = \mathrm{Col}(L_\Sigma)$.
 6:     Find $V \in \mathbb{R}^{k \times k}$ such that $V^\top (U^\top L_\Sigma U) V = I_{k \times k}$. Set $W = UV$.
 7:     Find a partition $\mathcal{P}$ of $[n]$ such that $|\mathcal{P}| = 3$ and $\mathrm{rank}(\Psi(W\theta)_{I,J}) = k$ for distinct $I, J \in \mathcal{P}$. Let
        L_Ψ and D_Ψ be the low-rank and the diagonal parts returned by SUBROUTINE($\Psi(W\theta), \mathcal{P}$).
 8:     Let Ω be the set of (left) singular vectors, with unique singular values, of $W^\top L_\Psi W$.
 9:     Let $S \in \mathbb{R}^{n \times k}$ be a matrix with columns $\{(W^+)^\top \omega : \omega \in \Omega\}$, where $W^+ = (W^\top W)^{-1} W^\top$.
10: **Part 3: Finding A and Λ.**
11:     Let $\widehat{A} = $ ALGORITHM 1(Σ).
12:     Let $\widehat{B}$ be a left inverse of $\widehat{A}$. Let $C = \widehat{B}S$.
13:     Reorder the rows and columns of C to make it lower triangular. Call it $\tilde{C}$.
14: **return** Columns of $\widehat{A}$ and $\widehat{\Lambda} = I - \mathrm{diag}(\tilde{C})\,\tilde{C}^{-1}$.

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

The rank condition for matrix A ensures the existence of a partition $\mathcal{P}$ of $[n]$, such that, $|\mathcal{P}| = 3$ and $A_I \in \mathbb{R}^{n \times k}$ has full column rank for all $I \in \mathcal{P}$. However, we are not provided with such a partition and therefore in ALGORITHM 1 and ALGORITHM 2 we need to search for $\mathcal{P}$. The complexity order of this searching step is n^k. Here, we show that under an *incoherence* assumption about A, a random partitioning of its rows into three groups has the desired property, with fixed positive probability.

Definition 3.10. *Let $A = USV^\top$ be a thin singular value decomposition of A, where $U \in \mathbb{R}^{n \times k}$ has orthonormal columns, $S = \mathrm{diag}(\sigma_1(A), \dots, \sigma_k(A))$, and $V \in \mathbb{R}^{k \times k}$ is orthogonal. Define the* incoherence number *of A as:*

$$c_A := \max_{j \in [n]} \left\{ \frac{n}{k} \|U^\top e_j\|_2^2 \right\}. \tag{12}$$

Lemma 3.11. *Fix $\ell \in [n]$, and consider ℓ random submatrices $A_1, A_2, \dots, A_\ell$ of A obtained by the following process: for each row of A, independently choose one of the ℓ submatrices uniformly at random, and put the row in that submatrix. Fix $\delta \in (0,1)$. Then,*

$$\mathbb{P}\left\{ \sigma_k(A_v) \geq \sigma_k(A)/(2\sqrt{\ell}), \forall v \in [\ell] \right\} \geq 1 - \delta, \tag{13}$$

provided that $c_A \leq \frac{9}{32} \cdot \frac{n}{k\ell \ln \frac{k\ell}{\delta}}$.

Lemma 3.11 is proved in Appendix E. Using this lemma with $\ell = 3$, we obtain the following. For $A \in \mathbb{R}^{n \times k}$ with full column rank and a random partitioning $\mathcal{P}$ of its rows into three groups, all the

submatrices A_I, $I \in \mathcal{P}$ are full rank with probability at least $1 - \delta$, provided that

$$c_A \leq \frac{3}{32} \cdot \frac{n}{k \ln \frac{3k}{\delta}}. \tag{14}$$

4 Proof of the theorems

4.1 Proof of Theorem 3.1

Observe that

$$\begin{aligned} \Sigma = \mathbb{E}[xx^\top] &= \mathbb{E}[(Ah + \varepsilon)(Ah + \varepsilon^\top)] \\ &= A\mathbb{E}[hh^\top]A^\top + \mathbb{E}[\varepsilon\varepsilon^\top]. \end{aligned} \tag{15}$$

Since the hidden variables are linearly independent, $\mathbb{E}[hh^\top]$ is full rank. Otherwise, $v^\top \mathbb{E}[hh^\top]v = 0$ for some non-zero vector v. This implies that $\mathbb{E}[\|h^\top v\|^2] = 0$ and so $h^\top v = 0$ which leads to a contradiction. Find a partition $\mathcal{P}$ of $[n]$, such that $|\mathcal{P}| = 3$, and $\operatorname{rank}(\Sigma_{I,J}) = k$ for all distinct $I, J \in \mathcal{P}$. (Note that $\operatorname{rank}(\Sigma_{I,J}) = \operatorname{rank}(A_I \mathbb{E}[hh^\top]A_J^\top)$ and by rank condition, there exists such a partition $\mathcal{P}$). We first show that SUBROUTINE$(\Sigma, \mathcal{P})$ returns $A\mathbb{E}[hh^\top]A^\top$ and the diagonal matrix $\mathbb{E}[\varepsilon\varepsilon^\top]$.

Lemma 4.1. *Let $C = AB^\top + D$, with $A, B \in \mathbb{R}^{n \times k}$ and $D \in \mathbb{R}^{n \times n}$ a diagonal matrix. Suppose that for a fixed partition $\mathcal{P}$ of $[n]$, with $|\mathcal{P}| = 3$, all the submatrices A_I and B_I have full column rank k, for all $I \in \mathcal{P}$. Then, SUBROUTINE(C) returns $AB^\top$ and D.*

The proof of Lemma 4.1 is deferred to Appendix B.

Given that $\mathbb{E}[hh^\top]$ and A have full column rank, we have $\operatorname{Col}(A) = \operatorname{Col}(A\mathbb{E}[hh^\top]A^\top)$. Let $\{u_1, \ldots, u_k\}$ be any basis of $\operatorname{Col}(A\mathbb{E}[hh^\top]A^\top)$ containing vectors with k smallest ℓ_0 norm. Since all the columns of A have at most $d_{\max}$ non-zero entries, we have $\max_{i \in [k]} \|u_i\|_0 \leq d_{\max}$, by choice of vectors u_i. Next we show that due to the graph expansion property (Condition 3) and the parameter genericity property (Condition 4), vectors u_i are (scaled) columns of A. Observe that any vector u_i can be represented by a linear combination of columns of A, say $u_i = Av$. If $\|v\|_0 \geq 2$, then

$$\|u_i\|_0 = \|Av\|_0 > |\mathrm{N}_A(\operatorname{supp}(v))| - |\operatorname{supp}(v)| \geq d_{\max},$$

where the first inequality follows from parameter genericity property and the second one follows from the expansion property. This leads to a contradiction. Therefore, $\|v\|_0 = 1$, and u_i is scaled version of a column of A. Since u_i are linearly independent, different u_i correspond to different columns of A. Let $\tilde{A} = [u_1|\cdots|u_k]$. Then, there exists a permutation matrix Π and a diagonal matrix Δ such that $\tilde{A} = A\Pi\Delta$. We recover the scaling matrix Δ using the fact that A is in canonical form.

Let $\tilde{B}$ be a left inverse of $\tilde{A}$. We have

$$\tilde{B}A\mathbb{E}[hh^\top]A^\top\tilde{B}^\top = \Delta^{-1}\Pi^{-1}\mathbb{E}[hh^\top]\Pi^{-\top}\Delta^{-\top}. \tag{16}$$

Consequently,

$$\operatorname{diag}(\tilde{B}A\mathbb{E}[hh^\top]A^\top\tilde{B}^\top) = \operatorname{diag}(\Delta^{-1}\Pi^{-1}\mathbb{E}[hh^\top]\Pi^{-\top}\Delta^{-\top}) = \Delta^{-2}, \tag{17}$$

where the last step follows from $\operatorname{diag}(\mathbb{E}[hh^\top]) = I_{k \times k}$ as A is in a canonical form. Finally,

$$\tilde{A}(\operatorname{diag}(\tilde{B}A\mathbb{E}[hh^\top]A^\top\tilde{B}^\top))^{1/2} = \tilde{A}\Delta^{-1} = A\Pi. \tag{18}$$

Therefore, we have identified all columns of A.

15

4.2 Proof of Theorem 3.2

Recall that $\Sigma = A\mathbb{E}[hh^\top]A^\top + \mathbb{E}[\varepsilon\varepsilon^\top]$. Using Lemma 4.1, SUBROUTINE$(\Sigma, \mathcal{P})$ (in the second step of ALGORITHM 1) returns the low-rank part $L = A\mathbb{E}[hh^\top]A^\top$ and the diagonal part $\mathbb{E}[\varepsilon\varepsilon^\top]$. The following lemma shows that vectors s_i, returned by the loop (steps $(3) - (5)$), are scaled multiples of the columns of A.

Lemma 4.2. *Let $A \in \mathbb{R}^{n \times k}$ be a given matrix with rank k, and let $L^{1/2} \in \mathbb{R}^{n \times k}$ be such that $L^{1/2} = AM$, for an invertible $M \in \mathbb{R}^{k \times k}$. (Equivalently $\mathrm{Col}(A) = \mathrm{Col}(L)$). Fix $i \in [n]$ and consider the following optimization problem:*

$$\min_w \quad \|L^{1/2}w\|_1 \qquad \text{subject to } (e_i^\top L^{1/2})w = 1. \tag{19}$$

Under the following conditions, $s_i = L^{1/2}w$ is a scaling of the $\pi_i(1)$-th column of A. (Recall that $\pi_i(1)$ is the index of the entry with maximum absolute value in the i-th row of A).

(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 $v \in \mathbb{R}^{|\mathrm{N}_i|-1}$.*

Proof (Lemma 4.2). Consider the following equivalent formulation of Problem (19) obtained by the change of variables $z = Mw$, $b^\top = (e_i^\top L^{1/2})M^{-1}$:

$$\min_z \quad \|Az\|_1 \qquad \text{subject to } b^\top z = 1. \tag{20}$$

Observe that $b^\top$ is the i-th row of A. Denote the solution to Problem (19) by z_*. We aim to prove that z_* is supported on $\{\pi_i(1)\}$. We prove the desired result in two steps:

Claim 4.3. *Under Condition (i), we have $\mathrm{supp}(z_*) \subseteq \mathrm{supp}(b)$.*

Claim 4.4. *Under Condition (i) − (ii), we have $\mathrm{supp}(z_*) = \{\pi_i(1)\}$.*

Proof (Claim 4.3). Notice that $b^\top = e_i^\top A$, and so $\mathrm{supp}(b) = \mathrm{N}_i$. Define $z_0 \in \mathbb{R}^k$ by $z_0(j) := z_*(j)$ for all $j \in \mathrm{supp}(b)$, and $z_0(j) := 0$ for all $j \notin \mathrm{supp}(b)$. Also, let $z_1 := z_* - z_0$. Therefore, z_0 is also a feasible solution to Problem (20), since $b^\top z_0 = b^\top z_*$.

If $z_1 \neq 0$, then

$$\begin{aligned}
\|Az_*\|_1 &= \|A_{\mathrm{N}_i^2,[k]}\, z_*\|_1 + \|A_{(\mathrm{N}_i^2)^c,[k]}\, z_*\|_1 \\
&= \|A_{\mathrm{N}_i^2,[k]}\, (z_0 + z_1)\|_1 + \|A_{(\mathrm{N}_i^2)^c,[k]}\, z_1\|_1 \\
&\geq \|A_{\mathrm{N}_i^2,[k]}z_0\|_1 - \|A_{\mathrm{N}_i^2,[k]}z_1\|_1 + \|A_{(\mathrm{N}_i^2)^c,[k]}\, z_1\|_1 \\
&= \|Az_0\|_1 - \|A_{\mathrm{N}_i^2,[k]}z_1\|_1 + \|A_{(\mathrm{N}_i^2)^c,[k]}\, z_1\|_1 \\
&> \|Az_0\|_1,
\end{aligned}$$

where the last inequality follows from Condition (i) and the fact $\mathrm{supp}(z_1) \subseteq (\mathrm{N}_i)^c$. Therefore, z_0 is a feasible solution with smaller objective value, which contradicts the optimality of z_*. Therefore we conclude that $z_1 = 0$, and hence $\mathrm{supp}(z_*) \subseteq \mathrm{supp}(b)$. $\qquad\square$

16

Proof (Claim 4.4). By Claim 4.3, $\text{supp}(z_*) \subseteq \text{supp}(b) = N_i$. To lighten the notation, let $j = \pi_i(1)$, and define $z_0 := (e_j^\top z_*)e_j$ and $z_1 := z_* - z_0$. Suppose for sake of contradiction that $z_1 \neq 0$. Since $b^\top z_* = 1$, we have $z_0 = ((1 - b^\top z_1)/b_j)\, e_j$. Therefore (using the triangle inequality twice),

$$
\begin{aligned}
\|Az_*\|_1 &= \|A_{N_j,[k]}z_*\|_1 + \|A_{(N_j)^c,[k]}z_*\|_1 \\
&= \|A_{N_j,[k]}(z_0 + z_1)\|_1 + \|A_{(N_j)^c,[k]}z_1\|_1 \\
&\geq \|A_{N_j,[k]}z_0\|_1 - \|A_{N_j,[k]}z_1\|_1 + \|A_{(N_j)^c,[k]}z_1\|_1 \\
&= \|A_{N_j,[k]}((1 - b^\top z_1)/b_j)\, e_j\|_1 - \|A_{N_j,[k]}z_1\|_1 + \|A_{(N_j)^c,[k]}z_1\|_1 \\
&\geq (1/|b_j|)\|A_{N_j,[k]}e_j\|_1 - |b^\top z_1/b_j|\|A_{N_j,[k]}e_j\|_1 - \|A_{N_j,[k]}z_1\|_1 + \|A_{(N_j)^c,[k]}z_1\|_1.
\end{aligned}
$$

Since $z_1(j) = 0$, we have $|b^\top z_1| \leq |b|_{\pi_i(2)}\|z\|_1$ by Hölder's inequality, and therefore,

$$
\frac{|b^\top z_1|}{|b_j|} \leq \frac{|b|_{\pi_i(2)}\|z_1\|_1}{|b|_j} \leq (1 - \gamma_i)\|z_1\|_1.
$$

Moreover, by Condition (ii) and the fact $\text{supp}(z_1) \subseteq N_i \setminus j$,

$$
\|A_{N_j^c,[k]}z_1\|_1 > \|A_{N_j,[k]}z_1\|_1 + (1 - \gamma_i)\|A_{N_j,j}\|_1\|z_1\|_1.
$$

Putting the last three displayed inequalities together gives

$$
\|Az_*\|_1 > (1/|b_j|)\|A_{N_j,[k]}e_j\|_1 = \|A(e_j/b_j)\|_1.
$$

Since e_j/b_j is a feasible solution, the above strict inequality contradicts the optimality of z_*. Therefore we conclude that $z_1 = 0$, and $z_* = z_0 = e_j/b_j$. $\qquad\square$

Notice that $s_i = L^{1/2}w = AMw = Az_*$ and since $\text{supp}(z_*) = \{\pi_i(1)\}$, s_i is a scaled multiple of the $\pi_i(1)$-th column of A. This completes the proof of Lemma 4.2. $\qquad\square$

Now, we are ready to prove the theorem.

Given that Conditions $(i) - (ii)$ hold for all $i \in [n]$, using Lemma 4.2, the set $\mathcal{S} = \{s_1, \ldots, s_n\}$ consists of scaled multiples of the columns of A. Moreover, since $[k] \subseteq \{\pi_1(1), \ldots, \pi_n(1)\}$, $\mathcal{S}$ contains a scaled multiple of each column of A. In the loop (steps $(6)-(10)$), we choose a linearly independent set $\{v_1, \ldots, v_k\} \subseteq \mathcal{S}$. These are the (scaled multiples of the) columns of A. Hence, letting $\tilde{A} = [v_1|\cdots|v_k]$, there exists a diagonal matrix $\Delta \in \mathbb{R}^{k \times k}$ and a permutation matrix $\pi \in \mathbb{R}^{k \times k}$ such that $\tilde{A} = A\Pi\Delta$. Let $\tilde{B}$ be a left inverse of $\tilde{A}$. We have

$$
\tilde{B}L\tilde{B}^\top = \tilde{B}A\mathbb{E}[hh^\top]A^\top\tilde{B}^\top = \Delta^{-1}\Pi^{-1}\mathbb{E}[hh^\top]\Pi^{-\top}\Delta^{-\top}. \tag{21}
$$

Consequently,

$$
\text{diag}(\tilde{B}L\tilde{B}^\top) = \text{diag}(\Delta^{-1}\Pi^{-1}\mathbb{E}[hh^\top]\Pi^{-\top}\Delta^{-\top}) = \Delta^{-2}, \tag{22}
$$

where the last step follows from $\text{diag}(\mathbb{E}[hh^\top]) = I_{k \times k}$ as A is in a canonical form. Finally,

$$
\tilde{A}(\text{diag}(\tilde{B}L\tilde{B}^\top))^{1/2} = \tilde{A}\Delta^{-1} = A\Pi. \tag{23}
$$

Therefore, ALGORITHM 1 returns all the columns of A.

4.3 Proof of Theorem 3.4

We identify the matrices A_i (up to permutation of their columns) in a sequential manner. Let h_{L_i} denote the vector formed by the hidden variables in level L_i, for $i \in [m-1]$. Also, let ε_{L_i} be the noise vector formed by the noise variables associated to the hidden nodes in layer L_i, for $i \in [m-1]$. Write

$$\Sigma = A_{m-1} \mathbb{E}[h_{L_{m-1}} h_{L_{m-1}}^\top] A_{m-1}^\top + \mathbb{E}[\varepsilon_{L_{m-1}} \varepsilon_{L_{m-1}}^\top]. \tag{24}$$

By applying Theorem 3.1, we can identify the columns of A_{m-1}. Equivalently, we recover $\widetilde{A}_{m-1} = A_{m-1} \Pi_{m-1}$ for some permutation matrix Π_{m-1}. Let $\widetilde{B}_{m-1}$ be a left inverse of $\widetilde{A}_{m-1}$. As demonstrated in the proof of Theorem 3.1, we can decompose Σ into its low-rank and diagonal parts. Therefore we have access to $A_{m-1} \mathbb{E}[h_{L_{m-1}} h_{L_{m-1}}^\top] A_{m-1}^\top$. Now, notice that

$$\widetilde{B}_{m-1} A_{m-1} \mathbb{E}[h_{L_{m-1}} h_{L_{m-1}}^\top] A_{m-1}^\top \widetilde{B}_{m-1}^\top = \Pi_{m-1}^{-1} \mathbb{E}[h_{L_{m-1}} h_{L_{m-1}}^\top] \Pi_{m-1}^{-\top}. \tag{25}$$

In words, we can recover the second order moment of the hidden variables in level L_{m-1}, up to a permutation of the nodes within this layer. Using the same technique sequentially, we can recover all the columns of A_i for $i \in [m-1]$ and thus the entire DAG is identifiable up to permutation of hidden nodes within each level.

4.4 Proof of Theorem 3.7

Let $\eta := (\eta_1, \ldots, \eta_k)$ and $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)$. Using the model description, we have

$$x = A(I - \Lambda)^{-1} \eta + \varepsilon. \tag{26}$$

Define $M := A(I - \Lambda)^{-1} \in \mathbb{R}^{n \times k}$. Then

$$\begin{aligned}
\Sigma &= \mathbb{E}[xx^\top] \\
&= \mathbb{E}[(M\eta + \varepsilon)(M\eta + \varepsilon)^\top] \\
&= M\mathbb{E}[\eta\eta^\top]M^\top + \mathbb{E}[\varepsilon\varepsilon^\top] \\
&= M \operatorname{diag}(\sigma_{\eta_1}^2, \ldots, \sigma_{\eta_k}^2)M^\top + \operatorname{diag}(\sigma_{\varepsilon_1}^2, \ldots, \sigma_{\varepsilon_n}^2).
\end{aligned} \tag{27}$$

Given that A satisfies the rank condition, it is immediate to see that $M \operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k}^2)$ also satisfies the rank condition. Therefore, applying Lemma 4.1, we can decompose Σ into its low-rank part (L_Σ) and its diagonal part (D_Σ), where

$$L_\Sigma = M \operatorname{diag}(\sigma_{\eta_1}^2, \ldots, \sigma_{\eta_k}^2)M^\top, \tag{28}$$

$$D_\Sigma = \operatorname{diag}(\sigma_{\varepsilon_1}^2, \ldots, \sigma_{\varepsilon_1}^2). \tag{29}$$

Since A has full column rank, $U^\top L_\Sigma U \in \mathbb{R}^{k \times k}$ also has full rank; hence, the whitening step (Part 2 in ALGORITHM 2) is possible. We have

$$I = W^\top L_\Sigma W = W^\top M \operatorname{diag}(\sigma_{\eta_1}^2, \ldots, \sigma_{\eta_k}^2)M^\top W.$$

Therefore, the matrix $N := W^\top M \operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k}) \in \mathbb{R}^{k \times k}$ is an orthogonal matrix.

Lemma 4.5. *We have*

$$\Psi(\zeta) = \text{diag}(\mu_{\varepsilon_1}\zeta_1, \ldots, \mu_{\varepsilon_n}\zeta_n) + M \, \text{diag}(\mu_{\eta_1}, \ldots, \mu_{\eta_k}) \, \text{diag}(M^\top\zeta)M^\top. \tag{30}$$

Lemma 4.5 is proved in Appendix C.

Applying Lemma 4.1 again, we decompose $\Psi(W\theta)$ into its diagonal and low-rank parts.

$$L_\Psi = M \, \text{diag}(\mu_{\eta_1}, \ldots, \mu_{\eta_k}) \, \text{diag}(M^\top W\theta)M^\top, \tag{31}$$

$$D_\Psi = \text{diag}(\mu_{\varepsilon_1}(W\theta)_1, \ldots, \mu_{\varepsilon_n}(W\theta)_n). \tag{32}$$

Now, observe that

$$\begin{aligned}
W^\top L_\Psi W = \\
W^\top M \, \text{diag}(\mu_{\eta_1}, \ldots, \mu_{\eta_k}) \, \text{diag}(M^\top W\theta)M^\top W = \\
N \, \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-1} \, \text{diag}(\mu_{\eta_1}, \ldots, \mu_{\eta_k}) \, \text{diag}(M^\top W\theta) \, \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-1} N^\top
\end{aligned} \tag{33}$$

Since N is an orthogonal matrix, the above is an SVD of $W^\top L_\Psi W$, and $N_1, \ldots, N_k$ are singular vectors, where N_i denotes the i-th column of N. Note that $N_i = \sigma_{\eta_i} W^\top M_i$ for $i \in [k]$.

A key observation is that an SVD uniquely determines all singular vectors (up to sign) which have distinct singular values. Following a similar approach to [4], we sample θ uniformly at random over the sphere in $\mathbb{R}^k$ to ensure that all the singular values of $W^\top L_\Psi W$ are distinct. Define

$$D := \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-1} \, \text{diag}(\mu_{\eta_1}, \ldots, \mu_{\eta_k}) \, \text{diag}(M^\top W\theta) \, \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-1}. \tag{34}$$

Note that the diagonal of the matrix D is the following vector:

$$\begin{aligned}
& \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-1} \, \text{diag}(\mu_{\eta_1}, \ldots, \mu_{\eta_k}) \, \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-1} M^\top W\theta \\
& = \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-1} \, \text{diag}(\mu_{\eta_1}, \ldots, \mu_{\eta_k}) \, \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})^{-2} N^\top\theta.
\end{aligned}$$

Since θ is sampled uniformly over the sphere, and N is a rotation matrix, the distribution of $N^\top\theta$ is also uniform over the sphere. Consequently, all the singular values of $W^\top L_\Psi W$ are non-zero and distinct. Therefore, the set Ω (in step (8) of the algorithm) is given by

$$\Omega = \{\sigma_{\eta_i} W^\top M_i\}_{i=1}^k.$$

The columns of matrix S, defined in step (9) of the algorithm, are then

$$\begin{aligned}
\{(W^+)^\top\omega : \omega \in \Omega\} &= \{W(W^\top W)^{-1}\sigma_{\eta_i} W^\top M_i\}_{i=1}^k \\
&= \{W(W^\top W)^{-1}W^\top \sigma_{\eta_i} M_i\}_{i=1}^k = \{\sigma_{\eta_i} M_i\}_{i=1}^k,
\end{aligned}$$

where the last step holds since $W(W^\top W)^{-1}W^\top$ is a projection and $\text{Range}(W) = \text{Range}(U) = \text{Range}(L_\Sigma) = \text{Range}(M)$. Hence, there exists permutation Π_1, such that

$$S = M \, \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})\Pi_1 = A(I - \Lambda)^{-1} \, \text{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})\Pi_1.$$

Note that $\text{Col}(S) = \text{Col}(A)$ and by (ii). As demonstrated in the proof of Theorem 3.1, we can identify all the columns of A, as A satisfies the graph expansion and the parameter genericity property. Moreover, under the assumptions of Theorem 3.2, ALGORITHM 1(Σ) returns all columns

of A. Therefore, we can recover $\widehat{A} = A\Pi_2$, for a permutation matrix $\Pi_2 \in \mathbb{R}^{k \times k}$. Let $\widehat{B}$ be a left inverse of $\widehat{A}$. Then

$$C := \widehat{B}S = \widehat{B}A(I - \Lambda)^{-1}\operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})\Pi_1 = \Pi_2^{-1}(I - \Lambda)^{-1}\operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})\Pi_1.$$

Consider a topological ordering of the induced DAG on the hidden nodes. In such an ordering, for every directed edge (j, i), we have $j < i$. Hence, Λ would be a lower triangular matrix in a topological ordering. We proceed by reordering the rows and the columns of C to get a lower triangular matrix. This may be done in many different ways but we show that all possible permutations that make C lower triangular correspond to different topological orderings of the same DAG. Therefore, we can choose any such permuted version of C, call it $\tilde{C}$. Then there exists a topological ordering with corresponding matrix Λ, such that, $(I - \Lambda)^{-1}\operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k}) = \tilde{C}$ and thus $\Lambda = I - \operatorname{diag}(\tilde{C})\tilde{C}^{-1}$.

Let R_1 denote the set of rows in C with exactly one non-zero entry. In any lower triangular version of C, the rows in R_1 should appear on top. Furthermore, their non-zero entries should appear in the first R_1 columns. Note that rows in R_1 correspond to hidden nodes with no parent. Obviously, any ordering of them with labels $1, \ldots, |\mathsf{R}_1|$ is faithful to topological orderings. Now, we can remove these nodes from the DAG (equivalently eliminate the R_1 columns and rows from C) and repeat the same argument. Therefore, different permuted versions of C which are lower triangular correspond to different topological orderings of the DAG. This completes the proof.

Acknowledgements

References

[1] P. Abbeel, D. Koller, and A. Ng. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, 7:1743–1788, 2006.

[2] R. Ali, T. Richardson, P. Spirtes, and J. Zhang. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 21th Conference on Uncertainty in Artificial Intelligence*, 2005.

[3] A. Anandkumar, K. Chaudhuri, D. Hsu, S. M. Kakade, L. Song, and T. Zhang. Spectral methods for learning multivariate latent tree structure. In *Advances in Neural Information Processing Systems*, 2011.

[4] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y.-K. Liu. Two SVDs Suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation. arXiv:1204.6703v3, 2012.

[5] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky. High-dimensional structure learning of Ising models: local separation criterion. *Annals of Statistics*, 40(3):1346–1375, 2012.

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

[7] T. O. Awokuse and D. A. Bessler. Vector autoregressions, policy analysis, and directed acyclic graphs: An application to the U.S. economy. *Journal of Applied Economics*, VI:1–24, 2003.

[8] F. Bach and M. Jordan. Beyond independent components: trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233, 2003.

[9] R. Bagozzi. *Causal models in marketing.* Theories in marketing series. Wiley, New York, 1980.

[10] D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, pages 17–35, 2007.

[11] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

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

[13] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence*, 1996.

[14] G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: some observations and algorithms. In *Intl. workshop APPROX Approximation, Randomization and Combinatorial Optimization.* Springer, 2008.

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

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

[17] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2003.

[18] M. Choi, V. Tan, A. Anandkumar, and A. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

[19] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

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

[21] C. Daskalakis, E. Mossel, and S. Roch. Optimal phylogenetic reconstruction. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, 2006.

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

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

[24] T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

[25] A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

[26] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, 2009.

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

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

[29] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. In *Proc. of NIPS*, 2011.

[30] M. Kohn and C. Schooler. Job conditions and personality: A longitudinal assessment of their reciprocal effects. *American Journal of Sociology*, 87(6):1257–1286, 1982.

[31] D. Koller, N. Friedman, L. Getoor, and B. Taskar. Graphical models in a nutshell. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[32] D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, pages 302–313, 1997.

[33] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[34] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proc. of Intl. Conf. on Machine learning*, pages 577–584, 2006.

[35] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

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

[37] J. Pearl. Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27(2):226–284, 1998.

[38] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, England, 2000.

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

[40] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *27th Conference on Uncertainty in Artificial Intelligence*, 2011.

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

[42] S. Roch and S. Snir. Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. In *Proceedings of the 16th Annual international conference on Research in Computational Molecular Biology*, RECOMB'12, pages 224–238, 2012.

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

[44] S. Shimizu, P. O. Hoyer, A. Hyvärisen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[45] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

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

[47] P. Spirtes. Graphical models, causal inference, and econometric models . *Journal of Economic Methodology*, 12:1:1–33, 2005.

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

[49] F. Theis. Towards a general independent subspace analysis. In *Advances in Neural Information Processing Systems*, 2007.

[50] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[51] B. Wheaton. The sociogenesis of psychological disorder. *American Sociological Review*, 43:383–403, 1978.

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

A   Proof of Remark 2.5

Let $\tilde{M} := M + Z$. We first establish some definitions.

Definition A.1. *We call a vector* fully dense *if all of its entries are non-zero.*

Definition A.2. *We say a matrix has the* Null Space Property (NSP) *if its null space does not contain any fully dense vector.*

Claim A.3. *Fix any $S \subseteq [k]$ with $|S| \geq 2$, and set $R := \mathrm{N}_M(S)$. Let $\tilde{C}$ be a $|S| \times |S|$ submatrix of $\tilde{M}_{R,S}$. Then $\Pr(\tilde{C}$ has the NSP$) = 1$.*

Now, we are ready to prove Remark 2.5.

Proof (Remark 2.5). It follows from Claim A.3 that, with probability one, the following event holds: for every $S \subseteq [k]$ with $|S| \geq 2$, and every $|S| \times |S|$ submatrix $\tilde{C}$ of $\tilde{M}_{R,S}$, $\tilde{C}$ has the NSP. Henceforth condition on this event.

Now fix $v \in \mathbb{R}^k$ with $\|v\|_0 \geq 2$. Let $S := \text{supp}(v)$, $R := N_M(S)$ and $B := \tilde{M}_{R,S}$. Furthermore, let $u \in (\mathbb{R} \setminus \{0\})^{|S|}$ be the restriction of vector v to S; observe that u is fully dense. It is clear that $\|\tilde{M}v\|_0 = \|Bu\|_0$, so we need to show that

$$\|Bu\|_0 > |R| - |S|. \tag{35}$$

Suppose for sake of contradiction that Bu has at most $|R| - |S|$ non-zero entries. Then there is a subset of $|S|$ entries on which Bu is zero. This corresponds to a $|S| \times |S|$ submatrix of B which contains u in its null space, which means that this submatrix does not have the NSP—a contradiction. Therefore we conclude that Bu must have more than $|R| - |S|$ non-zero entries. $\square$

Proof (Claim A.3). Let $s = |S|$ and let $\tilde{C} = [\tilde{c}_1 | \tilde{c}_2 | \cdots | \tilde{c}_s]^\top$, where $\tilde{c}_i^\top$ is the i-th row of $\tilde{C}$. Also, let $C := [c_1 | c_2 | \cdots | c_s]^\top$ and $W := [w_1 | w_2 | \cdots | w_s]^\top$ be the corresponding submatrices of M and Z, respectively. For each $i \in [s]$, denote by $\mathcal{N}_i$ the null space of the matrix $\tilde{C}_i = [\tilde{c}_1 | \tilde{c}_2 | \cdots | \tilde{c}_i]^\top$. Finally let $\mathcal{N}_0 = \mathbb{R}^s$. Then, $\mathcal{N}_0 \supseteq \mathcal{N}_1 \supseteq \cdots \supseteq \mathcal{N}_s$. We need to show that, with probability one, $\mathcal{N}_s$ does not contain any fully dense vector.

If one of $\mathcal{N}_i$ does not contain any full dense vector then we are done. Suppose that $\mathcal{N}_i$ contains some fully dense vector v. Since C is a submatrix of $M_{R,S}$, every row $c_{i+1}^\top$ of C contains at least one non-zero entry. Therefore

$$v^\top \tilde{c}_{i+1} = \sum_{j \in [s]} v(j) \tilde{c}_{i+1}(j)$$
$$= \sum_{j \in [s]:c_{i+1}(j) \neq 0} v(j)(c_{i+1}(j) + w_{i+1}(j))$$

where $\{w_{i+1}(j) : j \in [s] \text{ s.t. } c_{i+1}(j) \neq 0\}$ are independent random variables (from Z). Moreover, they are of $\tilde{c}_1, \ldots, \tilde{c}_i$ and thus of v. By assumption on the distribution of the $w_{i+1}(j)$,

$$\mathbb{P}\left[v \in \mathcal{N}_{i+1} \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = \mathbb{P}\left[\sum_{j \in [s]:c_{i+1}(j) \neq 0} v(j)(c_{i+1}(j) + w_{i+1}(j)) = 0 \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 0. \tag{36}$$

Consequently,

$$\mathbb{P}\left[\dim(\mathcal{N}_{i+1}) < \dim(\mathcal{N}_i) \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 1 \tag{37}$$

for all $i = 0, \ldots, s - 1$. As a result, with probability one, $\dim(\mathcal{N}_s) = 0$. $\square$

B Proof of Lemma 4.1

For each $I \in \mathcal{P}$, let $U_I, V_I \in \mathbb{R}^{|I| \times k}$ be any matrices such that $U_I^\top A_I$ and $V_I^\top B$ are invertible. Then for any distinct $I, J, K \in \mathcal{P}$,

$$
\begin{aligned}
A_I B_I^\top &= A_I (B_J^\top V_J)(B_J^\top V_J)^{-1}(U_K^\top A_K)^{-1}(U_K^\top A_K)B_I^\top \\
&= A_I B_J^\top V_J (U_K^\top A_K B_J^\top V_J)^{-1} U_K^\top A_K B_I^\top.
\end{aligned}
\tag{38}
$$

Notice that for any distinct $I, J \in \mathcal{P}$, $C_{I,J} = A_I B_J^\top$. Since A_I and B_J have rank k, so does $C_{I,J}$. Let $U_I \in \mathbb{R}^{|I| \times k}$ and $V_J \in \mathbb{R}^{|J| \times k}$ be respectively the matrices of left and right singular vectors of $C_{I,J}$ (corresponding to non-zero singular values). Since U_I and A_I have the same range, it follows that $U_I^\top A_I$ is invertible. Similarly $V_J^\top B_J$ is invertible. Using identity (38), we obtain

$$
A_I B_I^\top = C_{I,J} V_J (U_K^\top \Sigma_{K,J} V_J)^{-1} U_K^\top C_{K,I},
\tag{39}
$$

for any distinct $I, J, K \in \mathcal{P}$. Therefore D can be determined as $D_{I,I} = C_{I,I} - A_I B_I^\top$ for $I \in \mathcal{P}$ and $L = AB^\top$ is subsequently determined as $L = C - D$.

C Proof of Lemma 4.5

$$
\begin{aligned}
\Psi(\zeta) = \mathbb{E}[xx^\top \langle \eta, x \rangle] &= \mathbb{E}[(M\eta + \varepsilon)(\eta^\top M^\top + \varepsilon^\top)\langle \eta, M\eta + \varepsilon \rangle] \\
&= \mathbb{E}[(M\eta\eta^\top M^\top + \varepsilon\varepsilon^\top + M\eta\varepsilon^\top + \varepsilon\eta^\top M^\top)(\varepsilon^\top + \eta^\top M^\top)\zeta] \\
&= \mathbb{E}[\varepsilon\varepsilon^\top \varepsilon^\top \zeta + M\eta\eta^\top M^\top (\eta^\top M^\top \zeta)] \\
&= \mathbb{E}[\varepsilon\varepsilon^\top \langle \varepsilon, \zeta \rangle] + M\mathbb{E}[\eta\eta^\top \langle \eta, M^\top \zeta \rangle]M^\top.
\end{aligned}
\tag{40}
$$

The proof is completed by showing that for any deterministic vector $v \in \mathbb{R}^k$, and any random vector $z = (z_1, \ldots, z_k)$ with zero mean uncorrelated entries, we have

$$
\mathbb{E}[zz^\top \langle z, v \rangle] = \mathrm{diag}(v)\,\mathrm{diag}(\mu_{z_1}, \ldots, \mu_{z_n}).
\tag{41}
$$

We compute the diagonal and off-diagonal entries separately.

$$
\mathbb{E}[z_i z_i \langle v, z \rangle] = v_i \mathbb{E}[z_i^3] + \sum_{k \neq i} v_k \sigma_{z_i}^2 \mathbb{E}[z_k] = v_i \mu_{z_i}.
\tag{42}
$$

For $j \neq i$

$$
\mathbb{E}[z_i z_j \langle v, z \rangle] = \mathbb{E}\Big[z_i z_j \sum_k v_k z_k\Big] = v_i \sigma_{z,i}^2 \mathbb{E}[z_j] + v_j \sigma_{z,j}^2 \mathbb{E}[z_i] + \sum_{k \neq i,j} v_k \mathbb{E}[z_i]\mathbb{E}[z_j]\mathbb{E}[z_k] = 0.
\tag{43}
$$

D Proof of Remark 3.8

Write

$$
\Sigma = A\mathbb{E}[hh^\top]A^\top + \mathbb{E}[\varepsilon\varepsilon^\top].
\tag{44}
$$

By Theorem 3.1, we can identify the columns of A, *i.e.*, we can recover $\tilde{A} = A\Pi_1$ for some permutation matrix Π_1. Also, as demonstrated in the proof of Theorem 3.1, we can decompose Σ into its low-rank part $A\mathbb{E}[hh^\top]A^\top$ and diagonal part $\mathbb{E}[\varepsilon\varepsilon^\top]$. Let $\tilde{B} \in \mathbb{R}^{k \times n}$ be a left inverse of $\tilde{A}$. Then,

$$\tilde{B}A\mathbb{E}[hh^\top]A^\top\tilde{B}^\top = \Pi_1^{-1}\mathbb{E}[hh^\top]\Pi_1^{-\top}. \tag{45}$$

Therefore, we have the second order moment of the hidden nodes (in some ordering of the nodes). Now consider k hidden nodes corresponding to the row (and columns of) $\Pi_1^{-1}\mathbb{E}[hh^\top]\Pi_1^{-\top}$. Label these nodes with $1, \ldots, k$. Using the oracle we can find a permutation π_2 which puts the hidden nodes in a topological ordering. Let Π_2 be the corresponding permutation matrix to π_2. Then $\tilde{\Sigma} := \Pi_2\Pi_1^{-1}\mathbb{E}[hh^\top]\Pi_1^{-\top}\Pi_2^\top$ is the second order moment of the hidden nodes in some topological ordering. By definition of a topological ordering, it is immediate to see that the coefficient matrix Λ is lower triangular in a topological ordering of the hidden nodes. Therefore, we can write

$$\tilde{\Sigma} = (I - \Lambda)^{-1}\mathbb{E}[\eta\eta^\top](I - \Lambda)^{-\top}, \tag{46}$$

where η is the vector formed by the noise variables η_i (in the corresponding topological ordering) and $\Lambda \in \mathbb{R}^{k \times k}$ is a lower triangular matrix with all diagonal entries equal to zero. Therefore,

$$\tilde{\Sigma}^{1/2} = (I - \Lambda)^{-1}\operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})Q, \tag{47}$$

for some rotation $Q \in \mathbb{R}^{k \times k}$. Notice that $L := (I - \Lambda)^{-1}\operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})$ is a lower triangular matrix with diagonal entries $\sigma_{\eta_1}, \ldots, \sigma_{\eta_k}$ which are all positive. Hence, using the LQ decomposition of $\tilde{\Sigma}^{1/2}$, we can recover L. (Recall that the LQ factorization is unique if we require that the diagonal entries of the lower triangular part are positive).

Finally, $\operatorname{diag}(L) = \operatorname{diag}((I - \Lambda)^{-1})\operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k}) = \operatorname{diag}(\sigma_{\eta_1}, \ldots, \sigma_{\eta_k})$. Therefore, $\Lambda = I - \operatorname{diag}(L)L^{-1}$. The result follows.

E   Proof of Lemma 3.11

Let $A = USV^\top$ be a thin singular value decomposition of A, where $U \in \mathbb{R}^{n \times k}$ has orthonormal columns, $S = \operatorname{diag}(\sigma_1(A), \ldots, \sigma_k(A))$, and $V \in \mathbb{R}^{k \times k}$ is an orthogonal matrix. Fix a partition index $v \in [\ell]$. Let $z_1, z_2, \ldots, z_n \in \{0, 1\}$ be independent indicator random variables such that $z_i = 1$ iff row i of A is included in A_v. Note that

$$
\begin{aligned}
A_v^\top A_v &= A^\top \operatorname{diag}(z_1, z_2, \ldots, z_n)A \\
&= \sum_{i=1}^n z_i A^\top e_i e_i^\top A = VS(\sum_{i=1}^n z_i U^\top e_i e_i^\top U)SV^\top.
\end{aligned}
\tag{48}
$$

Therefore

$$\sigma_k(A_v)^2 = \lambda_{\min}(A_v^\top A_v) \geq \lambda_{\min}(S)^2 \cdot \lambda_{\min}(\sum_{i=1}^n z_i U^\top e_i e_i^\top U) = \sigma_k(A)^2 \cdot \lambda_{\min}(\sum_{i=1}^n X_i), \tag{49}$$

where $X_i := z_i U^\top e_i e_i^\top U \in \mathbb{R}^{k \times k}$. Notice that $0 \preceq X_i$ and

$$\lambda_{\max}(X_i) \leq \|U^\top e_i\|_2^2 \leq \frac{k}{n}c_A. \tag{50}$$

Moreover,

$$\sum_{i=1}^{n} \mathbb{E}X_i = \sum_{i=1}^{n} \mathbb{P}(z_i = 1)U^\top e_i e_i^\top U = \frac{1}{\ell}U^\top U = \frac{1}{\ell}I. \tag{51}$$

By Lemma E.1,

$$\mathbb{P}\left\{\lambda_{\min}(\sum_{i=1}^{d} X_i) \leq \frac{1}{4\ell}\right\} \leq k \cdot e^{-(3/4)^2/(2\ell c_A k/n)} \leq \delta/\ell, \tag{52}$$

where the last inequality follows from the assumption on c_A. Therefore by Eq. (49), $\sigma_k(A_v) \geq \sigma_k(A)/(2\sqrt{\ell})$, with probability at least $1 - \delta/\ell$. A union bound over all $v \in [\ell]$ completes the proof.

Lemma E.1 (Matrix Chernoff bound [50]). *Consider a finite sequence $\{X_i\}$ of independent and symmetric $k \times k$ random matrices such that $0 \preceq X_i$ and $\lambda_{\max}(X_i) \leq r$ almost surely. Define $\mu_{\min} := \lambda_{\min}(\sum_i \mathbb{E}X_i)$. For any $\epsilon \in [0, 1]$, we have*

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_i X_i\right) \leq (1 - \epsilon)\mu_{\min}\right\} \leq k \cdot e^{-\epsilon^2 \mu_{\min}/(2r)}. \tag{53}$$