UC Irvine
UC Irvine Previously Published Works

Title

Learning Topic Models and Latent Bayesian Networks Under Expansion Constraints

Permalink

https://escholarship.org/uc/item/38c8m71n

Authors

Anandkumar, Animashree
Hsu, Daniel
Javanmard, Adel
et al.

Publication Date

2012-09-24

Peer reviewed

Learning Topic Models and Latent Bayesian Networks Under Expansion Constraints

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

May 27, 2013

Abstract

Unsupervised estimation of latent variable models is a fundamental problem central to numerous applications of machine learning and statistics. This work presents a principled approach for estimating broad classes of such models, including probabilistic topic models and latent linear Bayesian networks, using only second-order observed moments. The sufficient conditions for identifiability of these models are primarily based on weak expansion constraints on the topic-word matrix, for topic models, and on the directed acyclic graph, for Bayesian networks. Because no assumptions are made on the distribution among the latent variables, the approach can handle arbitrary correlations among the topics or latent factors. In addition, a tractable learning method via ℓ_1 optimization is proposed and studied in numerical experiments.

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) [49]. They have widespread applicability in artificial intelligence [19, 25, 41, 42], in the social sciences [13, 18, 40, 50, 51, 64], and as structural equation models in economics [12, 18, 33, 51, 60, 65].

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

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* on the DAG. Roughly speaking, expansion property states that every subset of hidden nodes has "enough" number of outgoing edges in the DAG, 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.

Another class of popular latent variable models are the probabilistic topic models [17]. In topic models, the latent variables correspond to the topics in a document which generate the (observed) words. Perhaps, the most widely employed topic model is the latent Dirichlet allocation (LDA) [16], which posits that the hidden topics are drawn from a Dirichlet distribution. Recent approaches have established that the LDA model can be learned efficiently using low-order (second and third) moments, using spectral techniques [4, 5]. The LDA model, however, cannot incorporate arbitrary correlations[1] among the latent topics, and various correlated topic models have demonstrated superior empirical performance, e.g. [15, 45], compared to LDA. However, learning correlated topic models is challenging, and further constraints need to be imposed to establish identifiability and provable learning.

A typical (exchangeable) topic model is parameterized by the topic-word matrix, i.e., the conditional distributions of the words given the topics, and the latent topic distribution, which determines the mixture of topics in a document. In this paper, we allow for arbitrary (non-degenerate) latent topic distributions, but impose expansion constraints on the topic-word matrix. In other words, the word support of different topics are not "too similar", which is a reasonable assumption. Thus, we establish expansion as an unifying criterion for guaranteed learning of both latent Bayesian networks and topic models.

1.1 Summary of contributions

We establish identifiability for different classes of topic models and latent Bayesian networks, and more generally, for linear latent models, and also propose efficient algorithms for the learning task.

1.1.1 Learning Topic Models

Learning under expansion conditions. We adopt a moment-based approach to learning topic models, and specifically, employ second-order observed moments, which can be efficiently estimated using a small number of samples. We establish identifiability of the topic models for arbitrary (non-degenerate) topic mixture distributions, under assumptions on the topic-word matrix. The support of the topic-word matrix is a bipartite graph which relates the topics to words. We impose a weak

---

[1]LDA models incorporate only "weak" correlations among topics, since the Dirichlet distribution can be expressed as the set of independently distributed Gamma random variables, normalized by their sum: if $y_i \sim \Gamma(\alpha_i, 1)$, we have $\left(\frac{y_1}{\sum_i y_i}, \frac{y_2}{\sum_i y_i}, \ldots\right) \sim \text{Dir}(\alpha)$.

(additive) expansion constraint on this bipartite graph. Specifically, let $A \in \mathbb{R}^{n \times k}$ denote the topic-word matrix, and for any subset of topics $S \subset [k]$ (*i.e.*, a subset of columns of A), let $\mathrm{N}(S)$ denote the set of neighboring words, *i.e.*, the set of words, the topics in S are supported on. We require that

$$|\mathrm{N}(S)| \geq |S| + d_{\max}, \tag{1}$$

where $d_{\max}$ is the maximum degree for any topic. Intuitively, our expansion property states that every subset of topics generates sufficient number of words. We establish that under the above expansion condition in (1), for generic[2] parameters (for non-zero entries of A), the columns of A are the sparsest vectors in the column span, and are therefore, identifiable.

In contrast, note that for all subsets of topics $S \subset [k]$, the condition $|\mathrm{N}(S)| \geq |S|$, is *necessary* for non-degeneracy of A, and therefore, for identifiability of the topic model from second order observed moments. This implies that our sufficient condition in (1) is close to the necessary condition for identifiability of sparse models, where the maximum degree of any topic $d_{\max}$ is small. Thus, we prove identifiability of topic models under nearly tight expansion conditions on the topic-word matrix. Since the columns of A are the sparsest vectors in the column span under (1), this also implies recovery of A through exhaustive search. In addition, we establish that the topic-word matrix can be learned efficiently through ℓ_1 optimization, under some (stronger) conditions on the non-zero entries of the topic-word matrix, in addition to the expansion condition in (1). We call our algorithm TWMLEARN as it learns the topic-word matrix.

Bayesian networks to model topic mixtures. The above framework does not impose any parametric assumption on the distribution of the topic mixture h (other than non-degeneracy), and employs second-order observed moments to learn the topic-word matrix A and the second-order moments of h. If h obeys a multivariate Gaussian distribution, then this completely characterizes the topic model. However, for general topic mixtures, this is not sufficient to characterize the distribution of h, and further assumptions need to be imposed. A natural framework for modeling topic dependencies is via Bayesian networks [43]. Moreover, incorporating Bayesian networks for topic modeling also leads to efficient approximate inference through belief propagation and their variants [63], which have shown good empirical performance on sparse graphs.

We consider the case where the latent topics can be modeled by a *linear* Bayesian network, and establish that such networks can be learned efficiently using second and third order observed moments through a combination of ℓ_1 optimization and spectral techniques. The proposed algorithm is called TMLEARN as it learns (correlated) topic models.

1.1.2 Learning (Single-View) Latent Linear Bayesian Networks

The above techniques for learning topic models are also applicable for learning latent linear models, which includes linear Bayesian networks discussed in the introduction. This is because our method relies on the presence of a linear map from hidden to observed variables. In case of the topic models, the topic-word matrix represents the linear map, while for linear Bayesian networks, the (weighted) DAG from hidden to observed variables is the linear map. Linear latent models are prevalent in a number of applications such as blind deconvolution of sound and images [44]. The popular independent component analysis (ICA) [37] is a special case of our framework, where the

---

[2]The precise definition for parameter genericity is given in Condition 3.

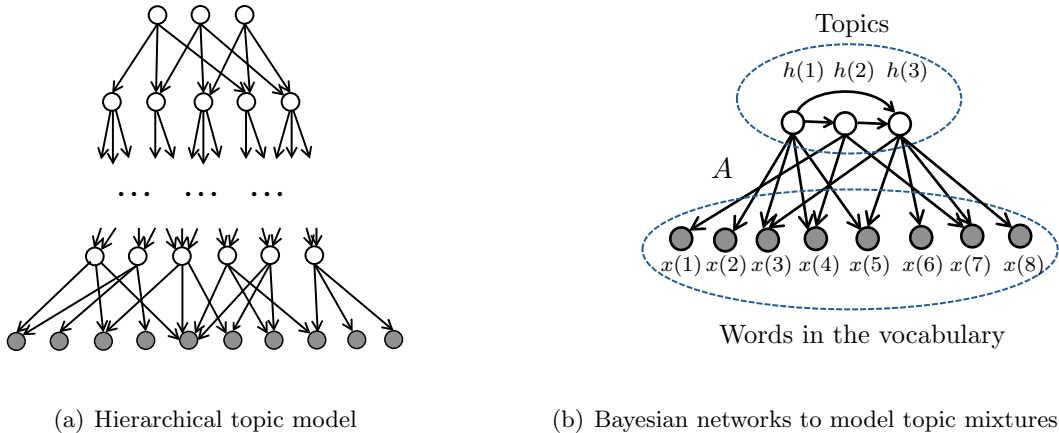(a) Hierarchical topic model   (b) Bayesian networks to model topic mixtures

Figure 1: Illustrations of hierarchical topic models and Bayesian networks for topic mixtures. Words and topics are respectively shown by shaded and white circles. Under the expansion property for the graph, we prove identifiability of these models from low order moments of the words.

sources (*i.e.*, the hidden variables) are assumed to be independent. In contrast, we allow for general latent distributions, and impose expansion conditions on the linear map from hidden to observed variables.

One key difference between topic models and other linear models (including linear Bayesian networks) is that topic models are multi-view (*i.e.*, have multiple words in the same document), while, for general linear models, multiple views may not be available. We require additional assumptions to provide recovery in the single-view setting. We prove recovery under certain rank conditions: we require that $n \geq 3k$, where n is the dimension of the observed random vector and k, the dimension of the latent vector, and the existence of a partition into three sets each with full column rank. Under these conditions, we propose simple matrix decomposition techniques to first "de-noise" the observed moments. These denoised moments are of the same form as the moments obtained from a topic model and thus, the techniques described for learning topic models can be applied on denoised moments. Thus, we provide a general framework for guaranteed learning of linear latent models under expansion conditions.

Hierarchical topic models. An important application of these techniques is in learning hierarchical linear models, where the developed method can be applied recursively, and the estimated second order moment of each layer can be employed to further learn the deeper layers. See Fig. 1(a) for an illustration.

Examples of graphs which can be learned. It is useful to consider some concrete examples which satisfy the expansion property in (1):

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 hierarchical models. We see that for $d \geq 2$, the model satisfies the expansion condition (1), but require $d \geq 3$ to satisfy the rank condition. See Fig. 2(a) for an illustration of a full ternary tree with latent variables.
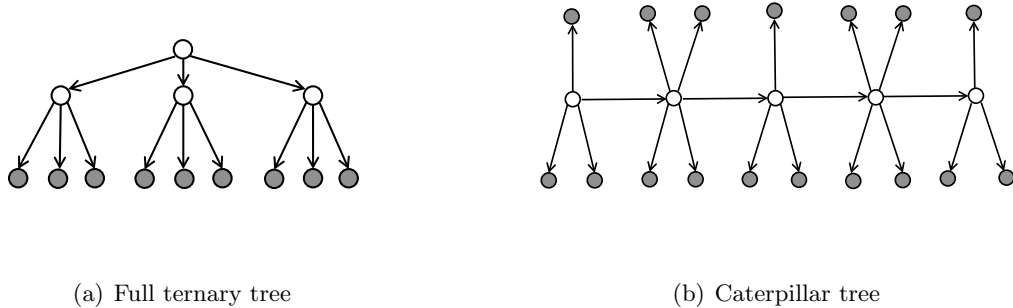
4

(a) Full ternary tree   (b) Caterpillar tree

Figure 2: Illustration of full ternary tree and caterpillar tree. Concrete examples of correlated topic models than can be learned using low order moments. Words and topics are respectively shown by shaded and white circles.

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 in (1).

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 $1 - \sqrt{1 - 2k/n} < \theta < 1/2$, *i.e.*, with probability converging to one as $n \to \infty$.

1.2 Our techniques

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

Dictionary learning and ℓ_1 optimization. We cast the topic models as *linear exchangeable multiview models* in Section 2.2 and demonstrate that the second order (cross) moment between any two words x_i, x_j satisfies

$$\mathbb{E}[x_i x_j^\top] = \mathbb{E}[\mathbb{E}[x_i x_j^\top | h]] = A\mathbb{E}[hh^\top]A^\top, \quad \forall\, i \neq j, \tag{2}$$

where $A \in \mathbb{R}^{n \times k}$ in the topic-word matrix, n is the vocabulary size, k is the number of topics, and h is the topic mixture. Thus, the problem of learning topic models using second order moments reduces to finding matrix A, given $A\mathbb{E}[hh^\top]A^\top$.

Indeed, further conditions need to be imposed for identifiability of A from $A\mathbb{E}[hh^\top]A^\top$. A natural non-degeneracy constraint is that the correlation matrix of the hidden topics $\mathbb{E}[hh^\top]$ be full rank, so that $\mathrm{Col}(A) = \mathrm{Col}(A\mathbb{E}[hh^\top]A^\top)$, where $\mathrm{Col}(\cdot)$ denotes the column span. Under the expansion

condition in (1), for generic parameters, we establish that the columns of A are the sparsest vectors in $\text{Col}(A)$, and are thus identifiable. To prove this claim, we leverage ideas from the work of Spielman et. al. [59], where the problem of sparsely used dictionaries is considered under probabilistic assumptions. In addition, we develop novel techniques to establish non-probabilistic counterpart of the result of [59]. A key ingredient in our proof is establishing that submatrices of the topic-word matrix, corresponding to any subset of columns and their neighboring rows, satisfy a certain null-space property under generic parameters and expansion condition in (1).

The above identifiability result implies recovery of the topic-word matrix A through exhaustive search for sparse vectors in $\text{Col}(A)$. Instead, we propose an efficient method to recover the columns of A through ℓ_1 optimization. We prove that ℓ_1 method recovers the matrix A, under the expansion condition in (1), and some additional conditions on the non-zero entries of A.

Spectral techniques for learning latent Bayesian networks. When the topic distribution is modeled via a linear Bayesian network, we exploit additional structure in the observed moments to learn the relationships among the topics, in addition to the topic-word matrix. Specifically, we assume that the topic variables obey the following linear equations:

$$h(j) = \sum_{\ell \in \mathsf{PA}_j} \lambda_{j\ell} h(\ell) + \eta(j), \quad \text{for } j \in [k], \tag{3}$$

where PA_j denotes the parents of node j in the directed acyclic graph (DAG) corresponding to the Bayesian network. Here, we assume that the noise variables $\eta(j)$ are non-Gaussian (*e.g.*, they have non-zero third moment or excess kurtosis), and are independent. We employ the ℓ_1 optimization framework discussed in the previous paragraph, and in addition, leverage the spectral methods of [4] for learning using second and third observed moments.

We first establish that the model in (3) reduces to independent component analysis (ICA), where the latent variables are independent components, and this problem can be solved via spectral approaches (*e.g.*, [4]). Specifically, denote $\Lambda = [\lambda_{i,j}]$, where $\lambda_{i,j}$ denotes the dependencies between different hidden topics in (3). Solving for the hidden topics h_j, we have $h = (I - \Lambda)^{-1}\eta$, where $\eta := (\eta(1), \ldots, \eta(k))$ denotes the independent noise variables in (3). Thus, the latent Bayesian network in (3) reduces to an ICA model, where $\eta := (\eta(1), \ldots, \eta(k))$ are the independent latent components, and the linear map from hidden to the observed variables is given by $A(I - \Lambda)^{-1}$, where A is the original topic-word matrix. We then apply spectral techniques from [4], termed as excess correlation analysis (ECA), 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 second SVD uses the third moment to find directions which exhibit information that is not captured by the second moment. Finally, in order to recover A from $A(I - \Lambda)^{-1}$, we exploit the expansion property in (1), and extract A as described previously through ℓ_1 optimization. The high-level idea is depicted in Fig. 3.

Matrix decomposition into diagonal and low-rank parts for general linear models. Our framework for learning topic models casts them as linear multiview models, where the words represent the multiple views of the hidden topic mixture h, and the conditional expectation of each word given the topic mixture h is a linear map of h. We extend our results for learning general linear models,
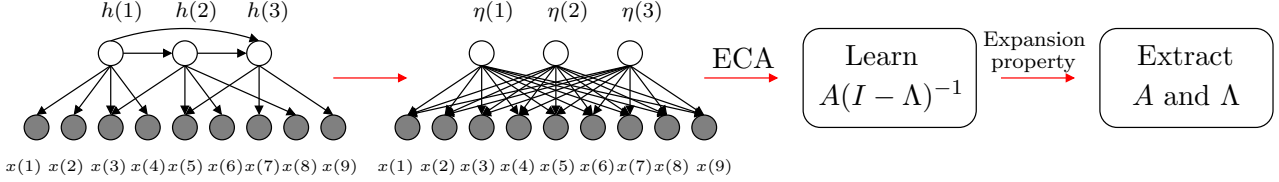
Figure 3: The high-level idea of the technique used for learning latent Bayesian networks. In the leftmost graph (original DAG) the hidden nodes depend on each other through the matrix Λ and the observed variables depend on the hidden nodes through the coefficient matrix A. We consider an equivalent DAG with new independent latent 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 $A(I - \Lambda)^{-1}$.

where such multiple views may not be available. Specifically, we consider

$$x(i) = \sum_{j \in \mathsf{PA}_i} a_{ij} h(j) + \varepsilon(i), \quad \text{for } i \in [n], \tag{4}$$

where $\{\varepsilon(i)\}_{i \in [n]}$ are uncorrelated and are independent from the hidden variables $\{h(j)\}_{j \in [k]}$. In this case, the second order moments $\Sigma := \mathbb{E}[xx^\top]$ satisfies

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

and has another noise component $\mathbb{E}[\varepsilon\varepsilon^\top]$, when compared to the second-order (cross) moment for topic models in (2). Note that the rank of $A\mathbb{E}[hh^\top]A^\top$ is k (under non-degeneracy conditions), where k is the number of topics. Thus, when k is sufficiently small compared to n, we can view Σ as the sum of a low-rank matrix and a diagonal one. We prove that under the rank condition that[3] $n \geq 3k$ (and the existence of a partition of three sets of columns of A such that each set has full column rank), $\mathbb{E}[xx^\top]$ can be decomposed into its low-rank component $A\mathbb{E}[hh^\top]A^\top$ and its diagonal component $\mathbb{E}[\varepsilon\varepsilon^\top]$. Thus, we employ matrix decomposition techniques to "de-noise" the second order moment and recover $A\mathbb{E}[hh^\top]A^\top$ from Σ. From here on, we can apply the techniques described previously to recover A through ℓ_1 optimization. Thus, we develop novel techniques for learning general latent linear models under expansion conditions.

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. 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

---

[3]It should be noted that other matrix decomposition methods have been considered previously [22, 36, 56]. Using these techniques, we can relax Condition 5 to $k \leq n/2$, but only by imposing stronger incoherence conditions on the low-rank component.

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 for future study. See Section 6 for the performance of our proposed algorithms under finite number of samples.

1.3 Related work

Probabilistic topic models have received widespread attention in recent years; see [17] for an overview. However, till recently, most learning approaches do not have provable guarantees, and in practice Gibbs sampling or variational Bayes methods are used. Below, we provide an overview of learning approaches with theoretical guarantees.

Learning topic models through moment-based approaches. A series of recent works aim to learn topic models using low order moments (second and third) under parametric assumptions on the topic distribution, e.g. single-topic model [6] (each document consists of a single topic), latent Dirichlet allocation (LDA) [4], independent components analysis (ICA) [37] (the different components of h, i.e., h_i are independent), and so on; see [5] for an overview. A general framework based on tensor decomposition is given in [5] for a wide range of latent variable models, including LDA and single topic models, Gaussian mixtures, hidden Markov models (HMM), and so on. These approaches do not impose any constraints on the topic-word matrix A (other than non-degeneracy). In contrast, in this paper, we impose constraints on A, and allow for any general topic distribution. Furthermore, we specialize the results to parametric settings where the topic distribution is a Bayesian network, and for this sub-class, we use ideas from the method of moments (in particular, the excess correlation method (ECA) of [4]) in conjunction with ideas from sparse dictionary learning.

Learning topic models through non-negative matrix factorization. Another series of recent works by Arora et. al. [9, 10] employ a similar philosophy as this paper: they allow for general topic distributions, while constraining the topic-word matrix A. They employ approaches based on non-negative matrix factorization (NMF), and exploit the fact that A is non-negative (recall that A corresponds to conditional distributions). The approach and the assumptions are quite different from this work. They establish guaranteed learning under the assumption that every topic has an *anchor* word, i.e. the word is uniquely generated from the topic, and does not occur under any other topic (with reasonable probability). Note that the presence of anchor words implies expansion constraint: $|\mathrm{N}(S)| \geq |S|$ for all subsets S of topics, where $\mathrm{N}(S)$ is the set of neighboring words for topics in S. In contrast, our requirement for guaranteed learning is $|\mathrm{N}(S)| \geq |S| + d_{\max}$, where $d_{\max}$ is the maximum degree of any topic. Thus our requirement is comparable to $|\mathrm{N}(S)| \geq |S|$, when $d_{\max}$ is small, and our approach does not require presence of anchor word. Additionally, our approach does not assume that the topic-word matrix A is positive, which makes it applicable for more general linear models, e.g. when the variables are not discrete and matrix A corresponds to a general mixing matrix (note that for discrete variables, A corresponds to conditional distribution and is thus non-negative).

Dictionary learning. As discussed in Section 1.2, we use some of the the ideas developed in the context of sparsely used dictionary learning problem. The problem setup there is that one is given a matrix X and is asked to find a pair of matrices A and M so that $\|X - AM\|$ is small and also M is

sparse. Here, A is considered as the dictionary being used. Spielman et. al [59] study this problem assuming that A is a full rank square matrix and the observation X is noiseless, i.e., $X = AM$. In this scenario, the problem can be viewed as learning a matrix X from its row space knowing that X enjoys some sparsity structure. Stating the problem this way clearly describes the relation to our work, as we also need to recover the topic-word matrix A from its second-order moments $A\mathbb{E}[hh^\top]A^\top$, as explained in Section 1.2.

The results of [59] are obtained assuming that the entries of M are drawn i.i.d. from a Bernoulli-Gaussian distribution. The idea is then to seek the rows of X sequentially, by looking for the sparse vectors in $\text{Row}(Y)$. Leveraging similar ideas, we obtain non-probabilistic counterpart of the results, i.e., without assuming any parametric distribution on the topic-word matrix. These conditions turn out to be intuitive expansion conditions on the support of the topic-word matrix, assuming generic parameters. Our technical arguments to arrive at these results are different than the ones employed in [59], since we do not assume any parametric distribution, and its application to learning topic models is novel. Moreover, in fact, it can be shown that the considered probabilistic models considered our [59], satisfy the expansion property (1) almost surely, and are thus, special cases under our framework. Variants of the sparse dictionary learning problem of [59] have also been proposed [32, 66]. For a detailed discussion on other works dealing with dictionary learning, refer to [59].

Linear structural equations. In general, structural equation modeling (SEM) is defined by a collection of equations $z_i = f_i(z_{\mathsf{PA}_i}, \varepsilon_i)$, where z_i's are the variables associated to the nodes. Recently, there has been some progress on the identifiability problem of SEMs in the fully observed linear models [35, 52, 53, 57]. More specifically, it has been shown that for linear functions f_i and non-Gaussian noise, the underlying graph $\mathcal{G}$ is identifiable [57]. Moreover, if one restricts the functions to be additive in the noise term and excludes the linear Gaussian case (as well as a few other pathological function-noise combinations), the graph structure $\mathcal{G}$ is identifiable [35,53]. Peters et. al. [52] consider Gaussian SEMs with linear functions, and the normally distributed noise variables with the same variances and show that the graph structure $\mathcal{G}$ and the functions are identifiable. However, none of these works deal with latent variables, or address the issue of efficiently learning the models. In contrast, our work here can be viewed as a contribution to the problem of identifiability and learning of linear SEMs with latent variables.

Learning Bayesian networks and undirected graphical models. 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 [29, 55], economics [12, 18, 33, 65], sociology [13, 18, 40, 64], and computer vision [25, 42]. The learning task has two main ingredients: structure learning and parameter estimation.

Structure estimation of probabilistic graphical models has been extensively studied in the recent years. It is well known that maximum likelihood estimation in fully observed tree models is tractable [26]. However, for general models, maximum likelihood 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 [23] find the graph structure by optimizing a score (*e.g.*, Bayesian Independence Criterion) in a greedy manner. Local test approaches attempt to build the graph based on local statistical tests on the samples, both for
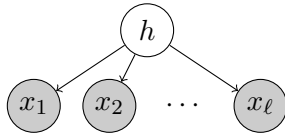
Figure 4: Exchangeable topic model with topic mixture h and x_i represents i-th word in the document.

directed and undirected graphical models [1, 7, 20, 34, 38, 61]. Convex relaxation approaches have also been considered for structure estimation (*e.g.*, [46, 54]).

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, 24, 27, 30]. Recently, approaches have been proposed for learning (undirected) latent graphical models with long cycles in certain parameter regimes [8]. In [21], latent Gaussian graphical models are estimated using convex relaxation approaches. The authors in [58] 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 Bayesian networks based on spectral techniques in conjunction with ℓ_1 optimization.

2 Model and sufficient conditions for identifiability

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 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 $\text{Row}(M)$ (similarly $\text{Col}(M)$) 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, $\text{diag}(v)$ is a diagonal matrix with the elements of v on the diagonal. For a matrix M, $\text{diag}(M)$ is a diagonal matrix with the same diagonal as M. Throughout $\otimes$ denotes the tensor product.

2.1 Overview of topic models

Consider the *bag-of-words* model for documents in which the sequence of observed words $x_1, x_2, \ldots, x_\ell$ in the document are *exchangeable*, i.e., the joint probability distribution is invariant to permutation of the indices. The well-known De Finetti's theorem [11] implies that such exchangeable models can be viewed as mixture models in which there is a latent variable h such that $x_1, x_2, \ldots, x_\ell$ are

conditionally i.i.d. given h and the conditional distributions are identical at all the nodes. See Fig.4 for an illustration.

In the context of document modeling, the latent variable h can be interpreted as a distribution over the topics occurring in a document. If the total number of topics is k, then h can be viewed as a distribution over the simplex Δ^{k-1}. The word generation process is thus a hierarchical process: for each document, a realization of h is drawn and it represents the proportion of topics in the documents, and for each word, first a topic is drawn from the topic mixture, and then the word is drawn given the topic.

Let $A = [a_{ij}] \in \mathbb{R}^{n \times k}$ denote the topic-word matrix, where $a_{i,j}$ denotes the conditional probability of word i occurring given that the topic j was drawn. It is convenient to represent the words in the document by n-dimensional random *vectors* $x_1, x_2, \ldots, x_\ell \in \mathbb{R}^n$. Specifically, we set

$$x_t = e_i \quad \text{if and only if} \quad \text{the } t\text{-th word in the document is } i, \quad t \in [\ell],$$

where $e_1, e_2, \ldots e_n$ is the standard coordinate basis for $\mathbb{R}^n$.

The above encoding allows for a convenient representation of topic models as linear models:

$$\mathbb{E}[x_i|h] = Ah, \quad \forall\, i \in [l],$$

and moreover the second order cross-moments (between two different words) have a simple form:

$$\mathbb{E}[x_i x_j^\top] = \mathbb{E}[\mathbb{E}[x_i x_j^\top |h]] = A\mathbb{E}[hh^\top]A^\top, \quad \forall\, i \neq j. \tag{5}$$

Thus, the above representation allows us to view topic models as linear models. Moreover, it allows us to incorporate other linear models, i.e. when x_i are not basis vectors. For instance, the independent components model is a popular framework, and can be viewed as a set of linear structural equations with latent variables. See Section 5 for a detailed discussion.

Thus, the learning task using second-order (exact) moments in (5) reduces to recovering A from $A\mathbb{E}[hh^\top]A^\top$, or equivalently $A\mathbb{E}[hh^\top]^{1/2}$.

2.2 Sufficient conditions for identifiability

We first start with some natural non-degeneracy conditions.

Condition 1 (Non-degeneracy). *The topic-word matrix $A := [a_{i,j}] \in \mathbb{R}^{n \times k}$ has full column rank and 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 such non-degeneracy assumptions, there is no hope of distinguishing different hidden nodes.

We now describe sufficient conditions under which the topic model becomes identifiable using second order observed moments. Given word observations $x_1, x_2, \ldots$, note that we can only hope to identify the columns of topic-word matrix A up to permutation because the model is unchanged if one permutes the hidden variable 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. (5) is unaltered if we both rescale all the coefficients $\{a_{ij}\}_{i \in [n]}$ and appropriately rescale the variable $h(j)$. Without further assumptions, we can only hope to recover a certain canonical form of A, defined as follows:

Definition 2.1. *We say A is in a* canonical form *if all of its columns have unit norm. In particular, the transformation $A \leftarrow A \operatorname{diag}(\|A_{[n],1}\|^{-1}, \|A_{[n],2}\|^{-1}, \ldots, \|A_{[n],k}\|^{-1})$ 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 its norm.

Thus, under the above non-degeneracy and scaling conditions, the task of recovering A from second-order (exact) moments in (5) reduces to recovering A from $\operatorname{Col}(A)$. Recall that our criterion for identifiability is that the sparsest vectors in the $\operatorname{Col}(A)$ correspond to the columns of A. We now provide sufficient conditions for this to occur, in terms of structural conditions on the support of A, and parameter conditions on the non-zero entries of A.

For structural conditions on the topic-word matrix A, we proceed by defining the *expansion property* of a graph which plays a key role in establishing our identifiability results.

Condition 2 (Graph expansion)**.** *Let $\mathcal{H}(\mathcal{V}_{\mathrm{hid}}, \mathcal{V}_{\mathrm{obs}})$ denote the bipartite graph formed by the support of A: $\mathcal{H}(i,j) = 1$ when $a_{i,j} \neq 0$, and 0 otherwise, and $\mathcal{V}_{\mathrm{hid}} := [k]$, $\mathcal{V}_{\mathrm{obs}} := [n]$. We assume that the $\mathcal{H}$ satisfies the following expansion property:*

$$|\mathrm{N}(S)| \geq |S| + d_{\max}, \quad \forall\, S \subset [k], |S| \geq 2, \tag{6}$$

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 $\mathcal{V}_{\mathrm{hid}}$.

Note that the condition $|\mathrm{N}(S)| \geq |S|$, for all subsets of hidden nodes $S \subset [k]$, is *necessary* for the matrix A to be full column rank. We observe that the above sufficient condition in (6) has an additional degree term $d_{\max}$, and is thus close to the necessary condition when $d_{\max}$ is small. Moreover, the above condition in (6) is only a weak additive expansion, in contrast to multiplicative expansion, which is typically required for various properties to hold, e.g. [14].

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

Condition 3 (Parameter genericity)**.** *We assume that the topic-word matrix A has the following parameter genericity property: for any $v \in \mathbb{R}^k$ with $\|v\|_0 \geq 2$, the following holds true.*

$$\|Av\|_0 > |\mathrm{N}_A(\operatorname{supp}(v))| - |\operatorname{supp}(v)|, \tag{7}$$

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

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

Remark 2.2. *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*

$$\mathbb{P}(M + Z \text{ does not satisfy Condition 3}) = 0. \tag{8}$$

Remark 2.2 is proved in Appendix B.

3 Identifiability result and Algorithm

In this section, we state our identifiability results and algorithms for learning the topic models under expansion conditions.

Theorem 3.1 (Identifiability of the Topic-Word Matrix). *Let* Pairs $:= \mathbb{E}[x_1 \otimes x_2]$ *be the pairwise correlation of the words. For the model described in Section 2.2 (Conditions 1, 2, 3), all columns of A are identifiable from* Pairs.

Theorem 3.1 is proved in Section A.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 $N_i := \{j \in [k] : A_{ij} \neq 0\}$ and $N_i^2 := \{l \in [n] : A_{lj} \neq 0 \text{ for some } j \in N_i\}$. Similarly, for $j \in [k]$, define $N_j := \{i \in [n] : A_{ij} \neq 0\}$ and $N_j^2 := \{l \in [k] : A_{il} \neq 0 \text{ for some } i \in N_j\}$. Thus, for a node i (either a topic or a word), N_i is the set of its neighbors and N_i^2 represents the set of nodes with distance exactly two from i. Therefore, if i is a word node, N_i^2 is the set of its siblings and if i is a topic word, N_i^2 is the set of topics with a common child. We further use superscript c to denote the set complement.

Theorem 3.2 (Recovery of the Topic-Word Matrix through ℓ_1-minimzation). *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* TWMLEARN *returns the columns of A in canonical form.*

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

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

Theorem 3.2 is proved in Section A.2. TWMLEARN is essentially the ER-SpUD presented in [59] for exact recovery of sparsely-used dictionaries, but the technical result and application in Theorem 3.2 are novel.

TWMLEARN involves solving n optimization problems and as the number of words becomes large, this requires a fast method to solve ℓ_1 minimization. Traditionally, the ℓ_1 minimization can be formulated as a linear programming (LP) problem. In particular, each of the ℓ_1 minimizations in TWMLEARN can be written as an LP with $2(n-1)$ inequality constraints and one equality constraint. However, the computational complexity of such a general-purpose formulation is often too high for large scale applications. Alternatively, one can use approximate methods which are significantly faster. There are several relevant algorithms with this theme, such as gradient projection [31, 39], iterative shrinkage-thresholding [28], and proximal gradient (Nestrov's method) [47, 48].

TWMLearn: Learning the topic-word matrix form pairwise correlations (Pairs).

Input: Pairwise correlation of the words (Pairs).
Output: Columns of A up to permutation.
 1: **for** each $i \in [n]$ **do**
 2:     Solve the optimization problem[4]

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

 3:     Set $s_i = \text{Pairs}^{1/2}w$, and let $\mathcal{S} = \{s_1, \ldots, s_n\}$.
 4: **for** each $j = 1, \ldots, k$ **do**
 5:     **repeat**
 6:         Let v_j be an arbitrary element in $\mathcal{S}$.
 7:         Set $\mathcal{S} = \mathcal{S}\backslash\{v_j\}$.
 8:     **until** $\text{rank}([v_1|\cdots|v_j]) = j$
 9: **return** $\widehat{A} = \left[\frac{v_1}{\|v_1\|}\Big|\cdots\Big|\frac{v_k}{\|v_k\|}\right].$

---

4   Bayesian networks for modeling topic distributions

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

Bayesian networks provide a natural framework for modeling topic dependencies, and we employ them here for modeling topic distributions. For these families, we prove identifiability and learning of the entire model, including the topic relationships and the topic-word matrix.

Bayesian networks, also known as belief networks, incorporate a set of causal and conditional independence through directed acyclic graphs (DAG) [49]. They have widespread applicability in artificial intelligence [19, 25, 41, 42], in the social sciences [13, 18, 40, 50, 51, 64], and as structural equation models in economics [12, 18, 33, 51, 60, 65].

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 k variables $h := (h(1), \ldots, h(k))$ that is Markov with respect to a DAG $\mathcal{G} = (\mathcal{H}, \mathcal{E})$ with $\mathcal{H} = \{1, \ldots, k\}$ [43]. More specifically, the joint probability $\mathbb{P}_\theta(h)$ factors as

$$\mathbb{P}_\theta(h) = \prod_{i=1}^{k} \mathbb{P}_\theta(h(i)|h_{\mathsf{PA}_i}), \tag{9}$$

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

We consider a subclass of DAG models for the topics in which the topics obey the linear relations

$$h(j) = \sum_{\ell \in \mathsf{PA}_j} \lambda_{j\ell} h(\ell) + \eta(j), \quad \text{for } j \in [k], \tag{10}$$

where $\eta(j)$ represents the noise variable at topic j. We further assume that the noise variables $\eta(j)$ are independent.

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. Without loss of generality, we assume that hidden (topic) variables $h(j)$, the observed (word) variables $x(i)$ and the noise terms $\varepsilon(i), \eta(j)$ are all zero mean. We also denote the variances of $\varepsilon(i)$ and $\eta(j)$ by $\sigma_{\varepsilon(i)}^2$ and $\sigma_{\eta(j)}^2$, respectively. Let $\mu_{\varepsilon(i)}$ and $\mu_{\eta(j)}$ respectively denote the third moment of $\varepsilon(i)$ and $\eta(j)$, i.e., $\mu_{\varepsilon(i)} := \mathbb{E}[\varepsilon(i)^3]$ and $\mu_{\eta(j)} := \mathbb{E}[\eta(j)^3]$. Define the skewness of $\eta(j)$ as:

$$\gamma_{\eta(j)} := \frac{\mu_{\eta(j)}}{\sigma_{\eta(j)}^3} \,. \tag{11}$$

Finally, define the following moments of the observed variables:

$$\begin{aligned} \text{Pairs} &:= \mathbb{E}[x_1 \otimes x_2], \\ \text{Triples} &:= \mathbb{E}[x_1 \otimes x_2 \otimes x_3]. \end{aligned} \tag{12}$$

It is convenient to consider the projection of Triples to a matrix as follows:

$$\text{Triples}(\zeta) := \mathbb{E}[x_1 \otimes x_2 \langle \zeta, x_3 \rangle],$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product.

Theorem 4.1. *Consider a DAG model which satisfies the model conditions described in Section 2.2 and the hidden variables are related through linear equations (10). If the noise variables $\eta(j)$ are independent and have non-zero skewness for $j \in [k]$, then the DAG model is identifiable from* Pairs *and* Triples(ζ), *for an appropriate choice of ζ. Furthermore, under the assumptions of Theorem 3.2,* TMLEARN *returns matrices A and Λ up to a permutation of hidden nodes.*

Theorem 4.1 is proven in Section A.3.

Notice that the only limitations on the noise variables $\eta(j)$ are that they are independent[5], and have non-zero skewness. Some common examples of non-zero skewness distributions are exponential, chi-squared and Poisson. Note that different topics may have different noise distributions.

Remark 4.2 (Special Cases). *A special case of the above result is when the DAG is empty, i.e. $\Lambda = 0$, and the topics $h(1), \ldots, h(k)$ are independent. This is popularly known as the independent components model (ICA), and similar spectral techniques have been proposed before for learning ICA [37]. Similarly, the ECA approach proposed above is also applicable for learning latent Dirichlet allocation (LDA), using suitably adjusted second and third order moments [4]. Note that for these special cases, we do not need to impose any constraints on the topic-word matrix A (other than non-degeneracy), since we can directly learn A and the topic distribution through ECA.*

Another immediate application of the technique used in the proof of Theorem 4.1 is in learning fully-observed linear Bayesian networks.

Remark 4.3 (Learning fully-observed BN's). *Consider an* arbitrary *fully-observed linear DAG:*

$$x(i) = \sum_{j \in \mathsf{PA}_i} \lambda_{ij} x(j) + \eta(i), \quad \text{for } i \in [n], \tag{13}$$

and suppose that the noise variables $\eta(i)$ have non-zero skewness. Then, applying the same argument as in the proof of Theorem 4.1, we can learn the matrix $(I - \Lambda)^{-1}$ (and hence Λ) from the second and third order moments (We have $A = I$ here).

TMLEARN: Learning topic models with correlated topics.

Input: Observable moments Pairs and Triples as defined in Eq. (12).
Output: Columns of A, matrix Λ (in a topological ordering).

1:   **Part 1: ECA.**
2:     Find a matrix $U \in \mathbb{R}^{n \times k}$ such that $\mathrm{Col}(U) = \mathrm{Col}(\mathrm{Pairs})$.
3:     Find $V \in \mathbb{R}^{k \times k}$ such that $V^\top (U^\top \mathrm{Pairs} U) V = I_{k \times k}$. Set $W = UV$.
4:     Let $\theta \in \mathbb{R}^k$ be chosen uniformly at random over the unit sphere.
5:     Let Ω be the set of (left) singular vectors, with unique singular values, of $W^\top \mathrm{Triples}(W\theta) W$.
6:     Let $S \in \mathbb{R}^{n \times k}$ be a matrix with columns $\{(W^+)^\top \omega : \omega \in \Omega\}$, where $W^+ = (W^\top W)^{-1} W^\top$.
7:   **Part 2: Finding A and Λ.**
8:     Let $\widehat{A} = \mathrm{TWMLEARN}(\mathrm{Pairs})$.
9:     Let $\widehat{B}$ be a left inverse of $\widehat{A}$. Let $C = \widehat{B} S$.
10:    Reorder the rows and columns of C to make it lower triangular. Call it $\tilde{C}$.
11: **return** Columns of $\widehat{A}$ and $\widehat{\Lambda} = I - \mathrm{diag}(\tilde{C}) \tilde{C}^{-1}$.

For sake of simplicity, TMLEARN is presented using the ECA method, which uses a single random direction θ and obtaining singular vectors of $W^\top \mathrm{Triples}(W\theta) W$. A more robust alternative to this, as described in [5], is to use the following power iteration to obtain the singular vectors $\{v_1, \ldots, v_k\}$; we use this variant in the simulations described in Section 6.

$\{v_1, \ldots, v_k\} \leftarrow$ random orthonormal basis for $\mathbb{R}^k$. Repeat:

1. For $i = 1, 2, \ldots, k$:

   - $v_i \leftarrow W^\top \mathrm{Triples}(W v_i) W v_i$.

2. Orthonormalize $\{v_1, \ldots, v_k\}$.

In principle, we can extend the above framework, combining spectral and ℓ_1 approaches, for learning other models on h. For instance, when the third order moments of h are sufficient statistics (e.g. when h is a graphical model with treewidth two), it suffices to learn the third order moments of h, i.e. $\mathbb{E}[h \otimes h \otimes h]$, where $\otimes$ denotes the outer product of vectors. This can be accomplished as follows: first employ ℓ_1 based approach to learn the topic-word matrix A, then consider the third order observed moments tensor $T := \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$. We have that

$$T(A^\dagger, A^\dagger, A^\dagger) = \mathbb{E}[h \otimes h \otimes h],$$

where $T(A^\dagger, A^\dagger, A^\dagger)$ denotes the multi-linear map of T under $A^\dagger$. For details on multi-linear transformation of tensors, see [5].

4.1 Learning using second-order moments

In Theorem 4.1, we prove identifiability and learning of hidden DAGs from second and third order observed moments. A natural question is what can be done if only the second order moment is

---

[5]We only require pairwise and triple-wise independence.

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 4.4. *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 a full column rank coefficient matrix A between the observed and hidden nodes. Further, 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* Pairs.

Remark 4.4 is proved in Appendix D.

5 Extension to general linear (single view) models

We have so far described a framework for identifiability and learning of topic models under expansion conditions. In fact, the developed framework holds for any linear *multi-view* model. Recall that if $x_1, x_2, \ldots$ are the words in the document, and h is the topic mixture variable, we have linearity $\mathbb{E}[x|h] = Ah$, and multiple (exchangeable and non-degenerate) views corresponding to different words in the document. In particular, the cross-moments between two different words x_1 and x_2, given h, is $\mathbb{E}[x_1 x_2^\top | h] = Ahh^\top A^\top$.

We now extend the results to a general framework where, unlike topic models, only a single observed view is available, and further assumptions are needed to learn in this setting.

Consider an observed random vector $x \in \mathbb{R}^n$ and a hidden random vector $h \in \mathbb{R}^k$. Let $\mathcal{G} = (\mathcal{V}_{\text{obs}} \cup \mathcal{V}_{\text{hid}}, \mathcal{E})$ denote the bipartite graph with observed nodes $\mathcal{V}_{\text{obs}} = \{x(1), \ldots, x(n)\}$ and hidden nodes $\mathcal{V}_{\text{hid}} = \{h(1), \ldots, h(k)\}$. Let $\varepsilon(i)$ be the noise variable associated with $x(i)$, for $i = 1, \ldots, n$ and denote the variance of $\varepsilon(i)$ 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 pairwise uncorrelated. The class of models considered are specified by the following assumptions.

Condition 4 (Linear model). *The observed and hidden variables obey the model*[6]

$$x(i) = \sum_{j \in \mathsf{PA}_i} a_{ij} h(j) + \varepsilon(i), \quad for \ i \in [n], \tag{14}$$

where $\{\varepsilon(i)\}_{i \in [n]}$ are pairwise uncorrelated and are independent from $\{h(j)\}_{j \in [k]}$. Furthermore, the matrix $A := [a_{i,j}] \in \mathbb{R}^{n \times k}$ has full column rank and 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 $i \in [k]$.*

Notice that the structure of $\mathcal{G}$ is defined by the non-zero coefficients in Eq. (14). 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*.

---

[6]Without loss of generality, assume that $x(i)$, $\varepsilon(i)$, $h(j)$ are all zero mean.

The above setting is prevalent in a number of applications such as the blind deconvolution of sound and images [44]. The independent component analysis (ICA) is a special case of the above setting, where the sources h_i are assumed to be independent. In contrast, in our setting, we allow for arbitrary distribution on h, and assume expansion (and rank) conditions on the coefficient matrix A.

Recall that in case of the topic models, A corresponds to the topic-word matrix. Moreover, in the topic model setting, no assumption is made on the noise variables ε, since the presence of cross-moments (between different words) enables us to remove the dependence on ε. However, in the single view case the second order observed moment $\Sigma := \mathbb{E}[xx^\top]$ is given by

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

We now discuss a rank condition on the coefficient matrix A, which allows us to remove the noise term $\mathbb{E}[\varepsilon\varepsilon^\top]$ from the second order moment Σ.

Condition 5 (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.

5.1 Matrix decomposition method for denoising

We now show that under the rank assumption in Condition 5, we can extract the noise terms ε from the observed moments through a matrix decomposition method.

Find a partition $\mathcal{P}$ of $[n]$, such that $|\mathcal{P}| = 3$, and $\mathrm{rank}(\Sigma_{I,J}) = k$ for all distinct $I, J \in \mathcal{P}$. (Note that $\mathrm{rank}(\Sigma_{I,J}) = \mathrm{rank}(A_I\mathbb{E}[hh^\top]A_J^\top)$ and by rank condition, there exists such a partition $\mathcal{P}$). We now show that the matrix decomposition procedure $\mathrm{DLD}(\Sigma, \mathcal{P})$ returns $A\mathbb{E}[hh^\top]A^\top$ and the diagonal matrix $\mathbb{E}[\varepsilon\varepsilon^\top]$.

Lemma 5.1. *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, $\mathrm{DLD}(C)$ returns $AB^\top$ and D.*

The proof of Lemma 5.1 is deferred to Appendix E.

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

The rank condition for matrix A in Condition 5 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. We now 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.

DLD: 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 $L = C - D$.

Definition 5.2. *Let $A = USV^\top$ be a thin singular value decomposition of A, where $U \in \mathbb{R}^{n \times k}$ has orthonormal columns, $S = \mathrm{diag}(\sigma_1(A), \ldots, \sigma_k(A))$, and $V \in \mathbb{R}^{k \times k}$ is orthogonal. Define the* incoherence number *of A as:*

$$c_A := \max_{j \in [n]} \left\{ \frac{n}{k} \| U^\top e_j \|_2^2 \right\}. \tag{15}$$

Lemma 5.3. *Fix $\ell \in [n]$, and consider ℓ random submatrices $A_1, A_2, \ldots, A_\ell$ 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,*

$$\mathbb{P}\left\{ \sigma_k(A_v) \geq \sigma_k(A)/(2\sqrt{\ell}), \forall v \in [\ell] \right\} \geq 1 - \delta, \tag{16}$$

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

Lemma 5.3 is proved in Appendix F. 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

$$c_A \leq \frac{3}{32} \cdot \frac{n}{k \ln \frac{3k}{\delta}}. \tag{17}$$

Thus, we have a procedure for denoising (i.e. recovering the noise terms ε) through random partitioning and matrix decomposition under appropriate rank condition. The coefficient matrix A can now be extracted from the denoised moments through the procedures listed in the previous sections, under expansion condition 2 and generic parameters condition 3 for the coefficient matrix A.

5.2 Application: learning hierarchical models

In the previous section, we developed a general framework for learning linear models with hidden variables.

We now apply the above results for learning hierarchical models, which consist of many layers of hidden variables. We first formally define hierarchical linear models.

Definition 5.4. *A hierarchical linear 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 hierarchical models. More specifically, consider a hierarchical 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 5) and parameter genericity property (Condition 3), and (bipartite) graph $\mathcal{G}_i$ has the expansion property (Condition 2).

Theorem 5.5. *Consider a hierarchical model 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 observed moment, i.e., $\Sigma = \mathbb{E}[xx^\top]$. Therefore, the entire model is identifiable up to permuting the nodes within each level.*

Theorem 5.5 is proved in Section A.4.

Remark 5.6. *By the definition of a hierarchical model, 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 5.5, we can still learn all the coefficient matrices A_i and the second order moment of the variables in layer L_1.*

6 Numerical experiments

In the previous sections, we proposed algorithms for learning topic models (multi-view), and general linear single view models. Our algorithms rely on low order (second and third order) moments of the observed variables. In presenting the results and the proofs we assumed that exact observed moments are available to emphasize the validity of the method. In general, these moments should be estimated from sampled data. This brings up the question of *sample complexity*, namely given a model $\mathcal{G}$, how many samples are required to estimate the model parameters with precision δ. We expect graceful sample complexity for the proposed algorithms as the low order moments can be reliably estimated from data. In this section, we consider two concrete examples of the *single view* linear models, and validate the performance of the proposed algorithms under finite number of samples.

The first example is a hierarchical model where we require the coefficient matrices between adjacent layers to be full rank. The second example is an illustration of a model in which the relations among the hidden nodes are described by a (general) DAG, and we require the coefficient matrix to be full rank.

Example 1. We validate our method on the following configuration.

- *Graph structure:* We consider a hierarchical model with three levels, L_1, L_2 and L_3. Levels L_1 and L_2 contain the hidden nodes with $n_1 = |L_1| = 5$, $n_2 = |L_2| = 30$ and level L_3 contains the observed nodes with $n_3 = |L_3| = 180$. Coefficient matrices $A_1 \in \mathbb{R}^{n_2 \times n_1}$ and $A_2 \in \mathbb{R}^{n_3 \times n_2}$,

respectively representing the linear relationships among the levels L_1, L_2 and the levels L_2, L_3, are constructed according to a Bernoulli-Gaussian model. More specifically, $A_1 = B \odot G$, where $B \in \mathbb{R}^{n_2 \times n_1}$ is an i.i.d. Bernoulli(p) matrix, and $G \in \mathbb{R}^{n_2 \times n_1}$ has i.i.d. standard normal entries. Further, $\odot$ indicates the entrywise product. In our experiment, we choose $p = 0.3$ to make the model satisfy the expansion property. Also recall that Theorem 3.2 assumes a positive gap γ_i between the maximum and the second maximum absolute values in the ith row, for $i \in [n]$. For the sake of simplicity, we consider the same gap γ for all the rows. More specifically, in each row of A_1 we change the entry with the maximum absolute value to ensure gap γ while keeping the sign of this entry unchanged. As we will see, γ has an important effect on sample complexity of the algorithm. A very small γ leads to a poor sample complexity and increasing γ improves the sample complexity of the algorithm. Similar model is used to generate A_2.

- *Noise variables:* For each noise variable, its variance is selected uniformly at random from the interval $[0.5, 1]$ and its distribution is chosen from a family of four distributions including (a) exponential; (b) poisson; (c) chi-squared; (d) gaussian. More specifically, for given variance σ^2, it is distributed as either $\mathsf{Exp}(\sigma^{-1})$, $\mathsf{Poisson}(\sigma^2)$, $(\sigma/\sqrt{2})\chi_1$, or $\mathsf{N}(0, \sigma^2)$ equally likely, where χ_1 denotes the chi-squared distribution with one degree of freedom.

In experiments we employed a slight variant of TWMLEARN to make it more robust to finite sample errors. This is essentially the same variant of ER-SpUD used in [59] (see ER-SpUD (proj)). For self-containedness, we present its details in Appendix G.

Following Lemma 5.3, we find partition $\mathcal{P}$ (the input of DLD) by randomly partitioning the rows of the corresponding coefficient matrix into three groups. With exact observed moments, any such partition leads to the decomposition of the corresponding matrix into its low rank and diagonal parts, with a fixed positive probability. However, with empirical moments, different partitions lead to different errors in estimating the coefficients. In experiments, we run DLD with 100 different partitions. Due to finite sample error, the retuned matrix D for each run is not necessarily a diagonal matrix. We compute the ratio of off-diagonal entries for each retuned D, i.e., $\sum_{i \neq j} |D_{ij}| / \sum_{i,j} |D_{ij}|$ and choose the partition $\mathcal{P}$ which leads to the minimum off-diagonal ratio.

We run TWMLEARN with empirical covariance $\widehat{\Sigma}$ to first learn A_1 and then A_2 as described in the proof of Theorem 5.5. More specifically, using n_{smp} independent realizations of the observed variables $x^{(1)}, \ldots, x^{(n_{\mathrm{smp}})} \in \mathbb{R}^{n_3}$, with $x^{(i)}$ representing the values of the observed nodes in the i-th realization, we let

$$m = \frac{1}{n_{\mathrm{smp}}} \sum_{i=1}^{n_{\mathrm{smp}}} x^{(i)}, \qquad \widehat{\Sigma} = \frac{1}{n_{\mathrm{smp}}} \sum_{i=1}^{n_{\mathrm{smp}}} (x^{(i)} - m)(x^{(i)} - m)^\top.$$

- *Measure of performance and the results:* Recall that coefficient matrices can be only specified up to permutation and scaling of their columns. In order to measure the algorithm performance on estimating a coefficient matrix $A \in \mathbb{R}^{n \times m}$, we define the following distance between A and the estimation $\widehat{A}$ returned by the algorithm.

$$\mathsf{dist}(A, \widehat{A}) = \frac{1}{\|A\|_F^2} \sum_{i=1}^{m} \min_{j \in [m], \nu} \|Ae_i - \nu \widehat{A}e_j\|^2$$

$$= \frac{1}{\|A\|_F^2} \sum_{i=1}^{m} \min_{j \in [m]} \|Ae_i - (e_i^\top A^\top \widehat{A}e_j)\widehat{A}e_j\|^2.$$

Here, the minimization over $j \in [m]$ is to remove the permutation ambiguity and the minimization over ν is to remove the scaling ambiguity. Further, since TWMLEARN returns the matrix in its canonical form, we have $\|\widehat{A}e_j\| = 1$ and the optimal ν is given by $\nu = (e_i^\top A^\top \widehat{A}e_j)$.

The support of coefficient matrix A corresponds to the edges in the corresponding graph and is of particular interest. We define precision and recall in characterizing the support of A as follows:

$$\text{precision}(A, \widehat{A}) = \frac{|\{(i,j) : A_{ij} \neq 0, \widehat{A}_{ij} \neq 0\}|}{|\{(i,j) : \widehat{A}_{ij} \neq 0\}|}, \quad \text{recall } (A, \widehat{A}) = \frac{|\{(i,j) : A_{ij} \neq 0, \widehat{A}_{ij} \neq 0\}|}{|\{(i,j) : A_{ij} \neq 0\}|}.$$

In words, *precision* is the fraction of retrieved edges that are truly an edge and *recall* is the fraction of true edges that are retrieved.

We summarize the results in Table 1 for different values of γ and n_{smp}.

(a) $\gamma = 0.3$

n_{smp}	25,000	50,000	100,000	200,000	400,000
dist$(A_1, \widehat{A}_1)$	0.9283	0.8029	0.7656	0.6939	0.4813
precision$(A_1, \widehat{A}_1)$	0.3120	0.3228	0.3231	0.3325	0.3333
recall$(A_1, \widehat{A}_1)$	0.8478	0.8913	0.9130	0.9130	0.9130
dist$(A_2, \widehat{A}_2)$	0.2674	0.1516	0.1466	0.1299	0.0943
precision$(A_2, \widehat{A}_2)$	0.3355	0.3402	0.3497	0.3530	0.3566
recall$(A_2, \widehat{A}_2)$	0.9389	0.9518	0.9526	0.9599	0.9697

(b) $\gamma = 0.5$

n_{smp}	25,000	50,000	100,000	200,000	400,000
dist$(A_1, \widehat{A}_1)$	0.5942	0.4016	0.3205	0.1187	0.0661
precision$(A_1, \widehat{A}_1)$	0.3462	0.3462	0.3538	0.3615	0.3769
recall$(A_1, \widehat{A}_1)$	0.8824	0.8824	0.9020	0.9216	0.9608
dist$(A_2, \widehat{A}_2)$	0.0731	0.0338	0.0157	0.0084	0.0048
precision$(A_2, \widehat{A}_2)$	0.3437	0.3497	0.3552	0.3558	0.3581
recall$(A_2, \widehat{A}_2)$	0.9477	0.9641	0.9793	0.9811	0.9872

Table 1: Example 1. Hierarchical (single-view) model with level sizes $n_1 = 5, n_2 = 30, n_3 = 180$, and 1694 number of edges.

.

The scatterplots in Fig. 5 depict the points $(\widehat{A}_{1,ij}, A_{1,ij})$ and $(\widehat{A}_{2,ij}, A_{2,ij})$ for different values of n_{smp} and $\gamma = 0.5$. As the number of samples increases, the observed moments are estimated more accurately and the scatter points concentrate around the line with slope one. Further, for each value of n_{smp}, the error in estimating A_1 is larger than the error in estimating A_2. The reason is that we first apply TWMLEARN(PROJ) to estimate the coefficient matrix A_2, and then use this estimation to learn the coefficient matrix A_1. In other words the induced model between the observed nodes (level L_3) and the hidden nodes (level L_2) is estimated more accurately than the induced model among the hidden nodes (levels L_1, L_2).
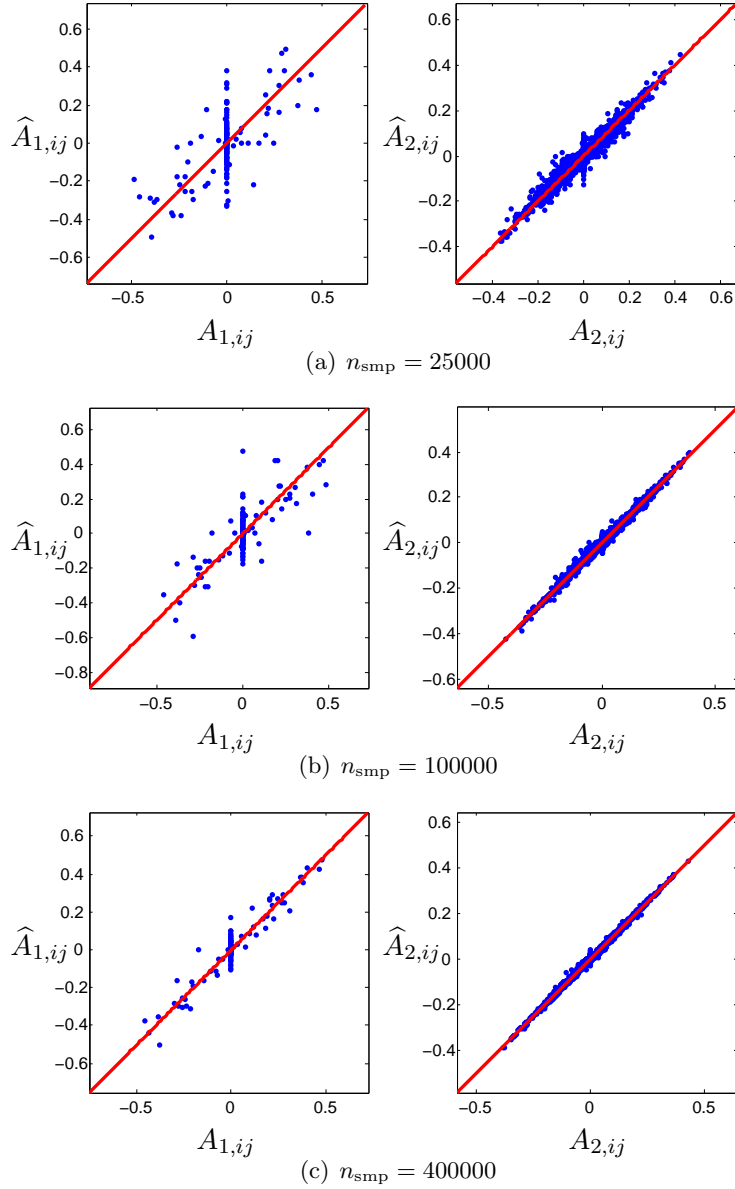
Figure 5: Scatterplots for Learning the hierarchical model in Example 1, using different values of n_{smp} and $\gamma = 0.5$.

Example 2. Our next example is a model in which the relationships among the hidden nodes are represented by a DAG model. The model contains $k = 25$ hidden nodes and $n = 150$ observed nodes. The linear relationships among the hidden nodes are described by a lower triangular coefficient matrix $\Lambda \in \mathbb{R}^{k \times k}$, which is chosen according to a Bernoulli-Gaussian model: The entries in the lower triangular part are non-zero with probability $p = 0.3$ and the values of the non-zero entries are chosen independently from standard normal distribution. The coefficient matrix $A \in \mathbb{R}^{n \times k}$, describing the relationships between the hidden nodes and the observed nodes, is constructed as per Bernoulli-

(a) $\gamma = 0.3$

n_{smp}	**200,000**	**300,000**	**400,000**	**500,000**
dist($\Lambda, \widehat{\Lambda}$)	0.7933	0.4627	0.3894	0.1778
precision($\Lambda, \widehat{\Lambda}$)	0.1168	0.1168	0.1168	0.1168
recall($\Lambda, \widehat{\Lambda}$)	1	1	1	1
dist($A, \widehat{A}$)	0.2818	0.2584	0.1894	0.0809
precision($A, \widehat{A}$)	0.2979	0.3248	0.3263	0.3337
recall($A, \widehat{A}$)	0.9391	0.9446	0.9492	0.9705

(b) $\gamma = 0.5$

n_{smp}	**200,000**	**300,000**	**400,000**	**500,000**
dist($\Lambda, \widehat{\Lambda}$)	0.4597	0.1820	0.0832	0.0492
precision($\Lambda, \widehat{\Lambda}$)	0.1168	0.1168	0.1168	0.1168
recall($\Lambda, \widehat{\Lambda}$)	1	1	1	1
dist($A, \widehat{A}$)	0.1777	0.0757	0.0478	0.0330
precision($A, \widehat{A}$)	0.3283	0.3302	0.3333	0.3352
recall($A, \widehat{A}$)	0.9548	0.9603	0.9695	0.9751

Table 2: Example 2. Bayesian network (single-view) model with $k = 25$ hidden nodes, $n = 150$ observed nodes, and 1177 number of edges.

.

Gaussian model in the previous experiment with $p = 0.3$, and then ensured to have gap γ between the maximum and the second maximum absolute values in each row.

Similar to the previous experiment, the noise variables have variances chosen uniformly at random from $[0.5, 1]$. Their distributions are chosen uniformly at random from a family of three distributions with non-zero skewness, namely (a) exponential; (b) poisson; (c) chi-squared.

In simulations, we used the power iteration to implement the ECA part as described in Section 4.

The results are summarized in Table 2. The scatterplots in Fig. 6 contains the points $(\widehat{\Lambda}_{ij}, \Lambda_{ij})$ and $(\widehat{A}_{ij}, A_{ij})$ for different values of n_{smp} and $\gamma = 0.5$.

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.
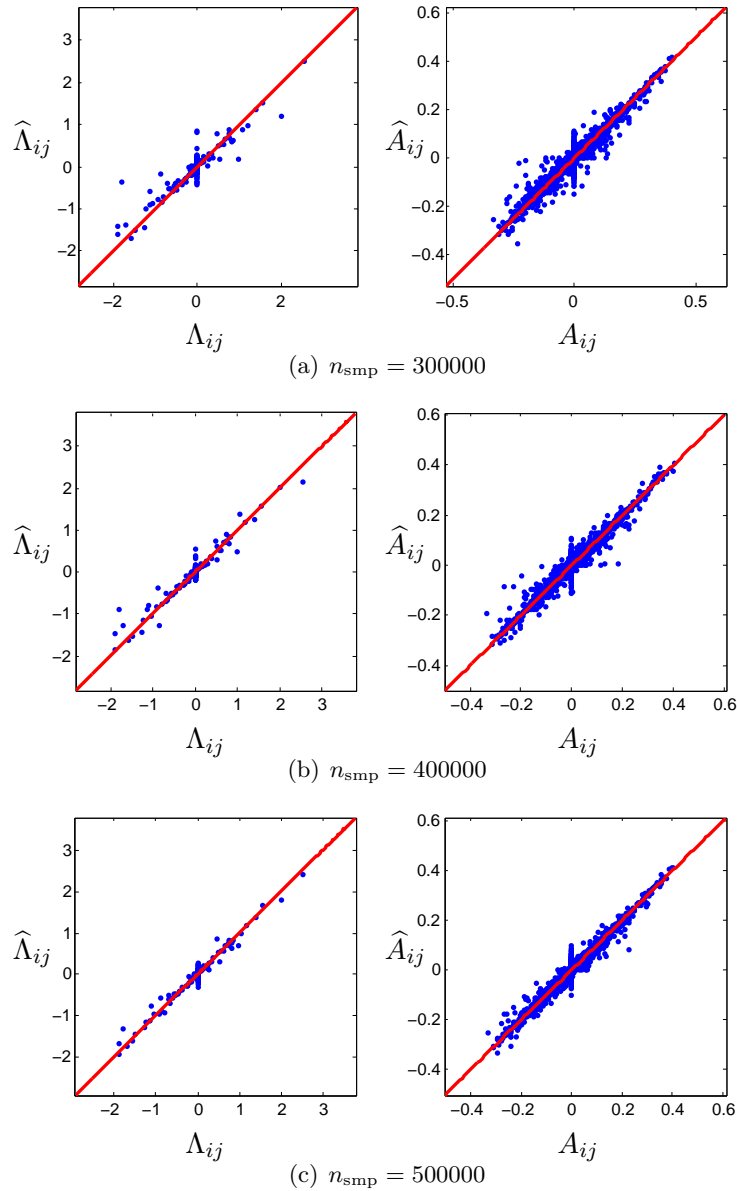
Figure 6: Scatterplots for Learning the model in Example 2, using different values of n_{smp} and $\gamma = 0.5$.

[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, R. Ge, D. Hsu, S. M. Kakade, and T. Telgarsky. Tensor decompositions for learning latent variable models. arXiv:1210.7559, 2012.

[6] A. Anandkumar, D. Hsu, and S. Kakade. A method of moments for mixture models and hidden Markov models. In *COLT*, 2012.

[7] 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.

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

[9] S. Arora, R. Ge, Y. Halpern, D. M. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. *ArXiv 1212.4777*, 2012.

[10] S. Arora, R. Ge, and A. Moitra. Learning topic models—going beyond svd. In *Symposium on Theory of Computing*, 2012.

[11] T. Austin. On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probab. Survey*, 5:80–145, 2008.

[12] 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.

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

[14] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 798–805. IEEE, 2008.

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

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

[17] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

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

[19] 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.

[20] 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.

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

[22] 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.

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

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

[25] 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.

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

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

[28] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

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

[30] 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.

[31] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

[32] L.-A. Gottlieb and T. Neylon. Matrix sparsification and the sparse null space problem. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 205–218, 2010.

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

[34] 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.

[35] 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.

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

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

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

[39] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An Interior-Point Method for Large-Scale L1-Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, Dec. 2007.

[40] 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.

[41] 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.

[42] 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.

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

[44] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1964–1971. IEEE, 2009.

[45] 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.

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

[47] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[48] Y. Nesterov. Gradient methods for minimizing composite objective function, 2007. ECORE Discussion Paper.

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

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

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

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

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

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

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

[63] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

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

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

[66] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.

A   Proof of the theorems

A.1   Proof of Theorem 3.1

Observe that

$$\text{Pairs} = \mathbb{E}[x_1 \otimes x_2] = \mathbb{E}[\mathbb{E}[x_1 \otimes x_2 | h]] = A\mathbb{E}[hh^\top]A^\top. \tag{18}$$

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.

Given that $\mathbb{E}[hh^\top]$ and A have full column rank, we have $\text{Col}(A) = \text{Col}(A\mathbb{E}[hh^\top]A^\top)$. Let $\{u_1, \ldots, u_k\}$ be any basis of $\text{Col}(A\mathbb{E}[hh^\top]A^\top)$ 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 2) and the parameter genericity property (Condition 3), 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

$$\|u_i\|_0 = \|Av\|_0 > |N_A(\text{supp}(v))| - |\text{supp}(v)| \geq d_{\max},$$

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 multiple of a column of A. Since $\{u_1, \ldots, u_k\}$ are linearly independent, different u_i's correspond to different columns of A and therefore columns of A, in a canonical form (up to sign), are given by $\{u_1/\|u_1\|, \ldots, u_k/\|u_k\|\}$.

A.2 Proof of Theorem 3.2

Recall that $\text{Pairs} = A\mathbb{E}[hh^\top]A^\top$. Using following lemma (with $L = \text{Pairs}^{1/2}$) shows that vectors s_i, returned by the first loop (steps $(1) - (3)$), are scaled multiples of the columns of A.

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

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

Under the following conditions, $s_i = Lw$ 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) $\|A_{(N_i^2)^c, (N_i)^c} v\|_1 > \|A_{N_i^2, (N_i)^c} v\|_1$ for all non-zero vectors $v \in \mathbb{R}^{|(N_i)^c|}$.

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

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

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

Observe that $b^\top$ is the i-th row of A. Denote the solution to Problem (20) by z_*. We aim to prove that z_* is supported on $\{\pi_i(1)\}$. We prove the desired result in two steps:

Claim A.2. *Under Condition (i), we have $\text{supp}(z_*) \subseteq \text{supp}(b)$.*

Claim A.3. *Under Condition (i) $-$ (ii), we have $\text{supp}(z_*) = \{\pi_i(1)\}$.*

Proof (Claim A.2). 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 $\mathrm{supp}(z_*) \subseteq \mathrm{supp}(b)$. $\qquad\square$

Proof (Claim A.3). By Claim A.2, $\mathrm{supp}(z_*) \subseteq \mathrm{supp}(b) = \mathrm{N}_i$. 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),

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

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 $\mathrm{supp}(z_1) \subseteq \mathrm{N}_i \setminus j$,

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

Putting the last three displayed inequalities together gives

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

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 $s_i = Lw = AMw = Az_*$ and since $\mathrm{supp}(z_*) = \{\pi_i(1)\}$, s_i is a scaled multiple of the $\pi_i(1)$-th column of A. This completes the proof of Lemma A.1. $\qquad\square$

Now, we are ready to prove the theorem.

Given that Conditions $(i) - (ii)$ hold for all $i \in [n]$, using Lemma A.1, 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 at least one scaled multiple of each column of A. In the second loop (steps $(4) - (8)$), 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 $\widehat{A} = [\frac{v_1}{\|v_1\|} | \cdots | \frac{v_k}{\|v_k\|}]$, there exists a permutation matrix Π, such that $\widehat{A}\Pi$ gives A in its canonical form (up to sign of each column).

A.3 Proof of Theorem 4.1

Let $\eta := (\eta(1), \ldots, \eta(k))$ and $\varepsilon := (\varepsilon(1), \ldots, \varepsilon(n))$. Using the model description, we have

$$\text{Pairs} = A\mathbb{E}[hh^\top]A^\top = A(I - \Lambda)^{-1}\mathbb{E}[\eta\eta^\top](I - \Lambda)^{-\top}A^\top . \tag{21}$$

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

$$\text{Pairs} = M\mathbb{E}[\eta\eta^\top]M^\top = M \operatorname{diag}(\sigma^2_{\eta(1)}, \ldots, \sigma^2_{\eta(k)})M^\top . \tag{22}$$

Since A has full column rank, $U^\top \text{Pairs} U \in \mathbb{R}^{k \times k}$ also has full rank; hence, the whitening step (Part 1 in TMLEARN) is possible. We have

$$I = W^\top \text{Pairs} W = W^\top M \operatorname{diag}(\sigma^2_{\eta(1)}, \ldots, \sigma^2_{\eta(k)})M^\top W.$$

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 A.4. *We have*

$$\text{Triples}(\zeta) = M \operatorname{diag}(\mu_{\eta(1)}, \ldots, \mu_{\eta(k)}) \operatorname{diag}(M^\top \zeta)M^\top . \tag{23}$$

Lemma A.4 is proved in Appendix C.

Now, observe that

$$
\begin{aligned}
&W^\top \text{Triples}(W\theta)W = \\
&W^\top M \operatorname{diag}(\mu_{\eta(1)}, \ldots, \mu_{\eta(k)}) \operatorname{diag}(M^\top W\theta)M^\top W = \\
&N \operatorname{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)})^{-1} \operatorname{diag}(\mu_{\eta(1)}, \ldots, \mu_{\eta(k)}) \operatorname{diag}(M^\top W\theta) \operatorname{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)})^{-1}N^\top
\end{aligned} \tag{24}
$$

Since N is an orthogonal matrix, the above is an SVD of $W^\top \text{Triples}(W\theta)W$, 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 $W^\top \text{Triples}(W\theta)W$ are distinct. Define

$$D := \operatorname{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)})^{-1} \operatorname{diag}(\mu_{\eta(1)}, \ldots, \mu_{\eta(k)}) \operatorname{diag}(M^\top W\theta) \operatorname{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)})^{-1}. \tag{25}$$

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

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

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 $W^\top\mathrm{Triples}(W\theta)W$ are non-zero and distinct. Therefore, the set Ω (in step (5) 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 (6) of the algorithm, are then

$$\{(W^+)^\top\omega : \omega \in \Omega\} = \{W(W^\top W)^{-1}\sigma_{\eta(i)}W^\top M_i\}_{i=1}^k$$
$$= \{W(W^\top W)^{-1}W^\top\sigma_{\eta(i)}M_i\}_{i=1}^k = \{\sigma_{\eta(i)}M_i\}_{i=1}^k,$$

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

$$S = M\,\mathrm{diag}(\sigma_{\eta(1)},\ldots,\sigma_{\eta(k)})\Pi_1 = A(I-\Lambda)^{-1}\,\mathrm{diag}(\sigma_{\eta(1)},\ldots,\sigma_{\eta(k)})\Pi_1.$$

Note that $\mathrm{Col}(S) = \mathrm{Col}(A)$. 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, TWMLEARN(Pairs) 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

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

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, $(I-\Lambda)^{-1}\,\mathrm{diag}(\sigma_{\eta(1)},\ldots,\sigma_{\eta(k)}) = \tilde{C}$ and thus $\Lambda = I - \mathrm{diag}(\tilde{C})\tilde{C}^{-1}$.

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.

A.4 Proof of Theorem 5.5

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 level L_i, for $i \in [m-1]$. Write

$$\Sigma = A_{m-1}\mathbb{E}[h_{L_{m-1}}h_{L_{m-1}}^\top]A_{m-1}^\top + \mathbb{E}[\varepsilon_{L_{m-1}}\varepsilon_{L_{m-1}}^\top]. \tag{26}$$

33

Applying Lemma 5.1, we can decompose Σ into its low-rank and diagonal parts. Therefore we have access to $A_{m-1}\mathbb{E}[h_{L_{m-1}}h_{L_{m-1}}^\top]A_{m-1}^\top$.

By a similar argument used in the proof of Theorem 3.1, we can identify the columns of A_{m-1}. Equivalently, we recover $\widehat{A}_{m-1} = A_{m-1}\Pi_{m-1}$ for some permutation matrix Π_{m-1}. Let $\widehat{B}_{m-1}$ be a left inverse of $\widehat{A}_{m-1}$. Now, notice that

$$\widehat{B}_{m-1}A_{m-1}\mathbb{E}[h_{L_{m-1}}h_{L_{m-1}}^\top]A_{m-1}^\top\widehat{B}_{m-1}^\top = \Pi_{m-1}^{-1}\mathbb{E}[h_{L_{m-1}}h_{L_{m-1}}^\top]\Pi_{m-1}^{-\top}. \tag{27}$$

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 level. Using the same technique sequentially, we can recover all the columns of A_i for $i \in [m-1]$ and thus the entire model is identifiable up to permutation of hidden nodes within each level.

B   Proof of Remark 2.2

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

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

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

Claim B.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 $\mathrm{Pr}(\tilde{C}$ has the NSP$) = 1$.*

Now, we are ready to prove Remark 2.2.

Proof (Remark 2.2). It follows from Claim B.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 $S := \mathrm{supp}(v)$, $R := \mathrm{N}_M(S)$ 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

$$\|Bu\|_0 > |R| - |S|. \tag{28}$$

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. $\qquad\square$

Proof (Claim B.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

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

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 $w_{i+1}(j)$,

$$
\mathbb{P}\left[v \in \mathcal{N}_{i+1} \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = \mathbb{P}\left[\sum_{j \in [s]: c_{i+1}(j) \neq 0} v(j)(c_{i+1}(j) + w_{i+1}(j)) = 0 \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 0. \quad (29)
$$

Consequently,

$$
\mathbb{P}\left[\dim(\mathcal{N}_{i+1}) < \dim(\mathcal{N}_i) \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 1 \quad (30)
$$

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

C Proof of Lemma A.4

$$
\begin{aligned}
\text{Triples}(\zeta) = \mathbb{E}[x_1 x_2^\top \langle \zeta, x_3 \rangle] &= \mathbb{E}[\mathbb{E}[x_1 x_2^\top \langle \zeta, x_3 \rangle | h]] \\
&= \mathbb{E}[\mathbb{E}[x_1|h]\mathbb{E}[x_2|h]^\top \langle \zeta, \mathbb{E}[x_3|h] \rangle] \\
&= \mathbb{E}[Ahh^\top A^\top \langle \zeta, Ah \rangle] \\
&= \mathbb{E}[M\eta\eta^\top M^\top \langle \zeta, M\eta \rangle] \\
&= M\mathbb{E}[\eta\eta^\top \langle \eta, M^\top \zeta \rangle]M^\top.
\end{aligned} \quad (31)
$$

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 independent entries, we have

$$
\mathbb{E}[zz^\top \langle z, v \rangle] = \text{diag}(v)\,\text{diag}(\mu_{z(1)}, \ldots, \mu_{z(n)}). \quad (32)
$$

We compute the diagonal and off-diagonal entries separately.

$$
\mathbb{E}[z(i)z(i)\langle v, z \rangle] = v(i)\mathbb{E}[z(i)^3] + \sum_{k \neq i} v(k)\sigma_{z(i)}^2 \mathbb{E}[z(k)] = v(i)\mu_{z(i)}. \quad (33)
$$

For $j \neq i$

$$
\begin{aligned}
\mathbb{E}[z(i)z(j)\langle v, z \rangle] &= \mathbb{E}\left[z(i)z(j)\sum_k v(k)z(k)\right] \\
&= v(i)\sigma_{z(i)}^2 \mathbb{E}[z(j)] + v(j)\sigma_{z(j)}^2 \mathbb{E}[z(i)] + \sum_{k \neq i,j} v(k)\mathbb{E}[z(i)]\mathbb{E}[z(j)]\mathbb{E}[z(k)] = 0.
\end{aligned}
$$

D Proof of Remark 4.4

Write

$$\text{Pairs} = A\mathbb{E}[hh^\top]A^\top. \tag{34}$$

By Theorem 3.1, we can identify the columns of A, *i.e.*, we can recover $\widehat{A} = A\Pi_1$ for some permutation matrix Π_1. Let $\widehat{B} \in \mathbb{R}^{k\times n}$ be a left inverse of $\widehat{A}$. Then,

$$\widehat{B}A\mathbb{E}[hh^\top]A^\top\widehat{B}^\top = \Pi_1^{-1}\mathbb{E}[hh^\top]\Pi_1^{-\top}. \tag{35}$$

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 $\widetilde{\text{Pairs}} := \Pi_2\Pi_1^{-1}\mathbb{E}[hh^\top]\Pi_1^{-\top}\Pi_2^\top$ 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

$$\widetilde{\text{Pairs}} = (I - \Lambda)^{-1}\mathbb{E}[\eta\eta^\top](I - \Lambda)^{-\top}, \tag{36}$$

where η is the vector formed by the noise variables $\eta(i)$ (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,

$$\widetilde{\text{Pairs}}^{1/2} = (I - \Lambda)^{-1}\text{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)})Q, \tag{37}$$

for some rotation $Q \in \mathbb{R}^{k\times k}$. Notice that $L := (I - \Lambda)^{-1}\text{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)})$ 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 $\widetilde{\text{Pairs}}^{1/2}$, 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, $\text{diag}(L) = \text{diag}((I - \Lambda)^{-1})\text{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)}) = \text{diag}(\sigma_{\eta(1)}, \ldots, \sigma_{\eta(k)})$. Therefore, $\Lambda = I - \text{diag}(L)L^{-1}$. The result follows.

E Proof of Lemma 5.1

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 $I, J, K \in \mathcal{P}$,

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

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

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

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 $L = C - D$.

F   Proof of Lemma 5.3

Let $A = USV^\top$ be a thin singular value decomposition of A, where $U \in \mathbb{R}^{n \times k}$ has orthonormal columns, $S = \mathrm{diag}(\sigma_1(A), \ldots, \sigma_k(A))$, 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

$$
\begin{aligned}
A_v^\top A_v &= A^\top \mathrm{diag}(z_1, z_2, \ldots, z_n) A \\
&= \sum_{i=1}^{n} z_i A^\top e_i e_i^\top A = VS(\sum_{i=1}^{n} z_i U^\top e_i e_i^\top U) SV^\top.
\end{aligned}
\tag{40}
$$

Therefore

$$
\sigma_k(A_v)^2 = \lambda_{\min}(A_v^\top A_v) \geq \lambda_{\min}(S)^2 \cdot \lambda_{\min}(\sum_{i=1}^{n} z_i U^\top e_i e_i^\top U) = \sigma_k(A)^2 \cdot \lambda_{\min}(\sum_{i=1}^{n} X_i),
\tag{41}
$$

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

$$
\lambda_{\max}(X_i) \leq \|U^\top e_i\|_2^2 \leq \frac{k}{n} c_A.
\tag{42}
$$

Moreover,

$$
\sum_{i=1}^{n} \mathbb{E} X_i = \sum_{i=1}^{n} \mathbb{P}(z_i = 1) U^\top e_i e_i^\top U = \frac{1}{\ell} U^\top U = \frac{1}{\ell} I.
\tag{43}
$$

By Lemma F.1,

$$
\mathbb{P}\left\{\lambda_{\min}(\sum_{i=1}^{d} X_i) \leq \frac{1}{4\ell}\right\} \leq k \cdot e^{-(3/4)^2/(2\ell c_A k/n)} \leq \delta/\ell,
\tag{44}
$$

where the last inequality follows from the assumption on c_A. Therefore by Eq. (41), $\sigma_k(A_v) \geq \sigma_k(A)/(2\sqrt{\ell})$, with probability at least $1 - \delta/\ell$. A union bound over all $v \in [\ell]$ completes the proof.

Lemma F.1 (Matrix Chernoff bound [62])**.** *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*

$$
\mathbb{P}\left\{\lambda_{\min}\left(\sum_i X_i\right) \leq (1 - \epsilon)\mu_{\min}\right\} \leq k \cdot e^{-\epsilon^2 \mu_{\min}/(2r)}.
\tag{45}
$$

G   TWMLearn(proj)

Below is the slight variant of TWMLearn used in numerical experiments.

TWMLearn(proj): Learning the topic-word matrix from pairwise correlations, using iterative projections.

Input: Second order moment of the observed variables (Pairs).
Output: Columns of A up to permutation.
1: Find a partition $\mathcal{P}$ of $[n]$ such that $|\mathcal{P}| = 3$ and $\text{rank}(\text{Pairs}_{I,J}) = k$ for distinct $I, J \in \mathcal{P}$.
2: Let L be the low-rank part returned by $\text{DLD}(\text{Pairs}, \mathcal{P})$.
3: Set $\mathcal{S} = \{0\} \subset \mathbb{R}^n$.
4: **for** each $i \in [k]$ **do**
5:     **for** each $j \in [n]$ **do**
6:         Solve the optimization problem

$$\min_w \ \|L^{1/2} w\|_1 \qquad \text{subject to } (e_j^\top L^{1/2}) P_{\mathcal{S}^\perp} w = 1.$$

   Denote the solution by w_{ij}.
7:     Set $w_i = \arg\min_{w_{i1}, \ldots, w_{in}} \|L^{1/2} w\|_0$, breaking ties arbitrarily.
8:     $\mathcal{S} = \mathcal{S} \cup \{w_i\}$.
9: **return** $\left\{ \frac{L^{1/2} w_i}{\|L^{1/2} w_i\|} \right\}_{i=1}^{k}$.