Learning Linear Bayesian Networks with Latent Variables

Anima Anandkumar

U.C. Irvine

Joint work with Daniel Hsu, Adel Javanmard, and Sham Kakade.
Latent Variable Modeling

Goal: Discover hidden effects from observed measurements

Document modeling
- Observed: words.
- Hidden: topics.

Social Network Modeling
- Observed: social interactions.
- Hidden: communities, relationships

Bio-Informatics
- Observed: gene expressions.
- Hidden: gene regulators.

Learning latent variable models: efficient methods and guarantees
Challenges and Approaches

Challenges: High-Dimensional Regime

- **Identifiability**: when can hidden variables be discovered?
- Design of learning algorithms with provable guarantees?
- Sample and Computational complexities?
Challenges and Approaches

Challenges: High-Dimensional Regime

- **Identifiability:** when can hidden variables be discovered?
- Design of learning algorithms with provable guarantees?
- Sample and Computational complexities?

**Our Approach: Two Perspectives**
Challenges and Approaches

Challenges: High-Dimensional Regime

- **Identifiability**: when can hidden variables be discovered?
- Design of learning algorithms with provable guarantees?
- Sample and Computational complexities?

**Our Approach: Two Perspectives**

Graphical Modeling

- **Bayesian networks**: Markov conditions on directed acyclic graphs.
Challenges and Approaches

Challenges: High-Dimensional Regime

- **Identifiability:** when can hidden variables be discovered?
- Design of learning algorithms with provable guarantees?
- Sample and Computational complexities?

Our Approach: Two Perspectives

Graphical Modeling

- **Bayesian networks:** Markov conditions on directed acyclic graphs.

Method of Moments

- **Linear models:** linear structural equation models (SEMs)
- Tractable approaches for solving equations (convex/non-convex).
Summary of Results

Model Class

- Linear Bayesian networks with hidden variables
- Multi-Level DAGs and DAGs with effective depth one.

Characterize Identifiability

- **Structural condition**: expansion of bipartite graph from hidden to observed nodes.
- **Parametric condition**: satisfied for generic parameters.

Learning Method

- Learning mixing matrix: from hidden to observed nodes.
  - Exploit sparsity in connections.
  - $\ell_1$ based method.
- Learning parameters in the hidden layer.
  - Exploit form of moments.
  - Spectral method.
Outline

1. Introduction
2. Model
3. Learning Algorithm
4. Conclusion
Linear Bayesian Networks

BN: Markov relationships on DAG

- $P_{\theta}(x) = \prod_{i=1}^{n} P_{\theta}(x_i | x_{Pa_i})$

Linear Model

- $n$ observed variables $\{x_i\}$ and $k$ hidden variables $\{h_i\}$.
- For each observed variable: $x_i = \sum_{j \in Pa_i} a_{ij} h_j + \varepsilon_i$.
- Condition on noise: Noise variables $\varepsilon_i$ are uncorrelated
- Non-degeneracy: Linear indep. on hidden variables, columns of $A$. 
Consider (exact) second-order observed moments

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

Learning

- In three stages: Denoising, unmixing and learning latent parameters
- **Denoising:** Separate noise \( \varepsilon \) from signal
- **Unmixing:** Separate mixing matrix \( A \) from hidden variables \( h_i \). Also known as blind deconvolution/dictionary learning.
- **Learning latent parameters:** learn deeper layers, learn hidden structures etc.
When $\varepsilon_i$ are uncorrelated, $\mathbb{E}[\varepsilon\varepsilon^\top]$ is a diagonal matrix.

Recall non-degeneracy conditions: $\text{Rank}(A\mathbb{E}[hh^\top]A^\top) = k$.

Thus, denoising is **Diagonal + Low Rank** when $n > k$, e.g. when $n > 3k$, can estimate diagonal part using off-diagonal parts.

For details, refer to the paper.
Denoising

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

- When \( \varepsilon_i \) are uncorrelated, \( \mathbb{E}[\varepsilon\varepsilon^\top] \) is a diagonal matrix.
- Recall non-degeneracy conditions: \( \text{Rank}(A\mathbb{E}[hh^\top]A^\top) = k \).
- Thus, denoising is **Diagonal + Low Rank** when \( n > k \), e.g. when \( n > 3k \), can estimate diagonal part using off-diagonal parts.
- For details, refer to the paper.

**Main focus:** unmixing \( A \) from \( A\mathbb{E}[hh^\top]A^\top \)
Some Intuitions on Blind Deconvolution

Main Task

Recover mixing matrix $A$ from $\mathbb{E}[hh^\top]A^\top$.

Ill-posed without further restrictions

One possibility: restriction on hidden variables $\{h_i\}$

- $\mathbb{E}[hh^\top]$ is diagonal: e.g. $h$ is the set of basis vectors in $\mathbb{R}^k$, when $h$ is uncorrelated, can obtain diagonal covariance matrix: (ICA), or when $h$ is drawn from Dirichlet distribution.
- No restrictions on $A$ (other than non-degeneracy).
- Recovery through third (or higher) order moment e.g. *simultaneous diagonalization*, through tensor decompositions (Anandkumar et. al. 2012).
Some Intuitions on Blind Deconvolution

Main Task

Recover mixing matrix $A$ from $A \mathbb{E}[hh^\top] A^\top$.

Ill-posed without further restrictions

One possibility: restriction on hidden variables $\{h_i\}$

- $\mathbb{E}[hh^\top]$ is diagonal: e.g. $h$ is the set of basis vectors in $\mathbb{R}^k$, when $h$ is uncorrelated, can obtain diagonal covariance matrix: (ICA), or when $h$ is drawn from Dirichlet distribution.
- No restrictions on $A$ (other than non-degeneracy).
- Recovery through third (or higher) order moment e.g. simultaneous diagonalization, through tensor decompositions (Anandkumar et. al. 2012).

Shortcoming: cannot handle arbitrary hidden dependencies.
Unmixing Task

Recover mixing matrix $A$ from $A \bar{\mathbf{E}} [h h^\top] A^\top$.

Different outlook: restriction on mixing matrix $A$

- No restrictions on hidden variables $\{h_i\}$ (other than non-degeneracy): can handle arbitrary hidden dependencies, e.g. correlated topic models.
- Restriction on support of $A$: corresponds to bipartite graph from hidden to observed layers.
- May be applicable in many settings, e.g. gene regulation, community memberships in social networks.
Unmixing Task

Recover mixing matrix $A$ from $A \mathbb{E} [hh^T] A^T$.

Different outlook: restriction on mixing matrix $A$

- No restrictions on hidden variables $\{h_i\}$ (other than non-degeneracy): can handle arbitrary hidden dependencies, e.g. correlated topic models.
- Restriction on support of $A$: corresponds to bipartite graph from hidden to observed layers.
- May be applicable in many settings, e.g. gene regulation, community memberships in social networks.
Sufficient Conditions for Identifiability

Unmixing Task: Recover $A$ from $A E [hh^\top] A^\top$

Structural Condition: (Additive) Graph Expansion

$|\mathcal{N}(S)| \geq |S| + d_{\text{max}}$, for all $S \subset [k]$

Parametric Conditions: Generic Parameters

$\|A v\|_0 > |\mathcal{N}_A(\text{supp}(v))| - |\text{supp}(v)|$

Identifiability Result

Under above conditions, $A$ can be uniquely recovered from $A E [hh^\top] A^\top$. 
Some Intuitions Behind Identifiability Result

- Identifiability of mixing matrix under graph expansion and for generic parameters.

**Intuitions**

- For non-degenerate $A \mathbb{E}[hh^\top] A^\top$, we know the $\text{Col}(A)$, the column space of $A$.
- Under above conditions, **sparsest vectors in $\text{Col}(A)$** are columns of $A$, and thus identifiable.

Unmixing: search for sparse vectors in $\text{Col}(A)$
Outline

1 Introduction
2 Model
3 Learning Algorithm
4 Conclusion
Tractable Algorithm for Unmixing

Unmixing Task

Recover mixing matrix $A$ from $A \mathbb{E}[hh^\top]A^\top$.

Exhaustive search

$$\min_{z \neq 0} \| Az \|_0$$

Convex relaxation

$$\min_z \| Az \|_1, \quad b^\top z = 1,$$
where $b$ is a row in $A$.

Change of Variables

$$\min_w \| (A \mathbb{E}[hh^\top]A^\top)^{1/2} w \|_1, \quad e_i^\top (A \mathbb{E}[hh^\top]A^\top)^{1/2} w = 1.$$  

Under “reasonable” conditions, the above program exactly recovers $A$.
Learning Latent Space Parameters

Recall so far..

Recover mixing matrix $A$ from $A \mathbb{E}[hh^\top] A^\top$.

Now learning hidden structures

- In general, $\mathbb{E}[hh^\top]$ is not enough to recover joint distribution of $h$

Learning Multi-level DAGs

Repeat this recursively, i.e., un-mix $\mathbb{E}[hh^\top]$ to recover higher layers.
Recall so far..

Recover mixing matrix $A$ from

$$A \mathbb{E}[hh^\top] A^\top.$$ 

Now learning hidden structures

- In general, $\mathbb{E}[hh^\top]$ is not enough to recover joint distribution of $h$

Learning Multi-level DAGs

Repeat this recursively, i.e., un-mix $\mathbb{E}[hh^\top]$ to recover higher layers.
Effective Depth

Each hidden variable is connected to at least one observed variable.

Linear Structural Equations

- Recall, \( x = Ah + \varepsilon \)
- Now additionally, \( h_j = \sum_{i \in \text{Pa}_j} \lambda_{ji} h_i + \eta_j \), or \( h = \Lambda h + \eta \)
- This implies that \( x = A(I - \Lambda)^{-1} \eta + \varepsilon \)
- \( \eta_i \) are uncorrelated: \( \mathbb{E}[\eta \eta^\top] \) is diagonal.

Spectral approach for learning
Effective Depth

Each hidden variable is connected to at least one observed variable.

Linear Structural Equations

- Recall, \( x = Ah + \varepsilon \)
- Now additionally, \( h_j = \sum_{i \in \text{Pa}_j} \lambda_{ji} h_i + \eta_j \), or \( h = \Lambda h + \eta \)
- This implies that \( x = A(I - \Lambda)^{-1} \eta + \varepsilon \)
- \( \eta_i \) are uncorrelated: \( \mathbb{E}[\eta \eta^\top] \) is diagonal.

Spectral approach for learning
Learning DAGs with Effective Depth

\[ x = A(I - \Lambda)^{-1}\eta + \varepsilon \]

- Employ spectral approach to learn \( A(I - \Lambda)^{-1} \).
- Therefore, \[ \mathbb{E}[xx^\top] = A(I - \Lambda)^{-1}\mathbb{E}[\eta\eta^\top](A(I - \Lambda)^{-1})^\top + \mathbb{E}[\varepsilon\varepsilon^\top] \]
- Similarly for third order moment, \[ \mathbb{E}[xx^\top\langle \eta, x \rangle] = \]
\[ A(I - \Lambda)^{-1}\mathbb{E}[\eta\eta^\top\langle \eta, A^\top\lambda \rangle](A(I - \Lambda)^{-1})^\top + \mathbb{E}[\varepsilon\varepsilon^\top\langle \lambda, \varepsilon \rangle] \]
- Simultaneous diagonalization of second and third order moments: through SVD or tensor decompositions.
- Un-mix \( A \) from \( A(I - \Lambda)^{-1} \) through \( \ell_1 \) optimization.

Learning both structure and parameters of depth-1 DAGs
Conclusion

Learning Linear Latent Bayesian Networks

- Considered learning with arbitrary hidden variable dependencies
- Constraints on the mixing matrix: expansion of bipartite graph from hidden to observed layer, generic parameters and non-degeneracy.
- Established identifiability of mixing matrix.
- Recovering mixing matrix through $\ell_1$ optimization.
- Able to learn multi-level DAGs and DAGs with effective depth 1

Outlook: Learning over-complete basis

- When more hidden variables than observed variables
- Require higher order moments
- Interesting questions on identifiability and efficient algorithms.