Learning Linear Bayesian Networks with Latent Variables

Adel Javanmard

Stanford University

joint work with Anima Anandkumar*, Daniel Hsu†, Sham Kakade†

* University of California, Irvine

† Microsoft Research, New England

Modern data

- Lots of high-dimensional data, but highly structured.
- Learning the underlying structure is central to:
  - Modeling
  - Dimensionality reduction/ summarizing data
  - Prediction

This talk:

Learning hidden (unobserved) variables that pervaded the data.

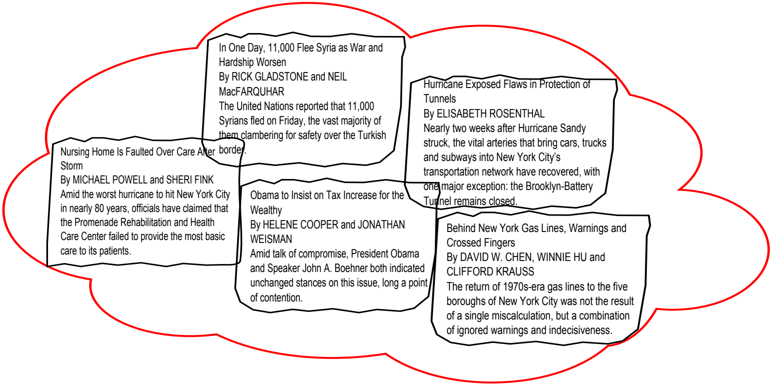
Modern data

- Lots of high-dimensional data, but highly structured.
- Learning the underlying structure is central to:
  - Modeling
  - Dimensionality reduction/ summarizing data
  - Prediction

This talk:

Learning hidden (unobserved) variables that pervaded the data.

Example: document modeling

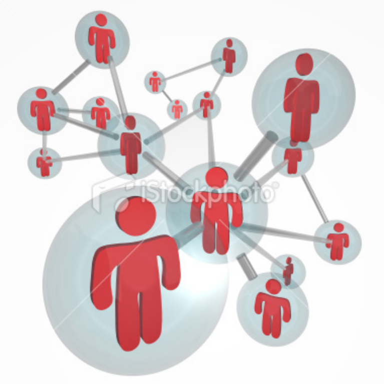


Observations: words Hidden variables: topics

Topics

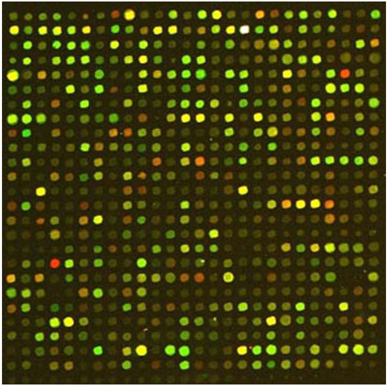
genome	disease	software
molecular	tuberculosis	system
sequence	penumonia	parallel
DNA	control	hardware
human	doctor	cyber
genetics	weak	network
map	resistance	data
project	fatal	program

Example: social network modeling



Observations: social interactions Hidden: communities, relationships

Example: bio-informatics

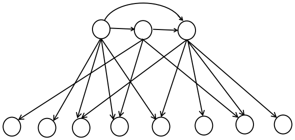


Observations: gene expressions   Hidden variables: gene regulators

Linear Bayesian Network

Markov relationship on DAG

- PA_i : parents of node i.
- $\mathbb{P}_\theta(z) = \prod_{i=1}^{n} \mathbb{P}_\theta(z_i | z_{\mathrm{PA}_i})$.



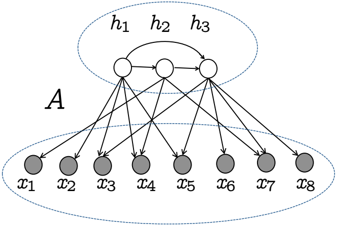
Linear model with latent nodes

- Observed variables $\{x_i\}$ and hidden variables $\{h_i\}$.
- Linear relations: $\boxed{x_i = \sum_{j \in \mathrm{PA}_i} a_{ij} h_j + \epsilon_i}$
- uncorrelated noise variables ϵ_i

Linear Bayesian Network

Markov relationship on DAG

- PA_i : parents of node i.
- $\mathbb{P}_\theta(z) = \prod_{i=1}^{n} \mathbb{P}_\theta(z_i | z_{\mathrm{PA}_i})$.



Linear model with latent nodes

- Observed variables $\{x_i\}$ and hidden variables $\{h_i\}$.
- Linear relations: $\boxed{x_i = \sum_{j \in \mathrm{PA}_i} a_{ij} h_j + \epsilon_i}$
- uncorrelated noise variables ϵ_i

Learning latent models

Goal: Given the observed data, learn structure and parameters of model.

Challenges:

- Identifiablity Many models can explain the observed data!
  - ICA: no edge between hidden nodes
  - LDA: hidden variables are drawn from a Dirichlet distribution
  - latent trees, graphical models with long cycles. [Anandkumar et.al. 2011, Choi et. al. 2011, Daskalakis et. al. 2006]
- Tractable learning algorithms:
  - Maximum likelihood (tractable on trees, NP-hard in general)
  - Expectation maximization [Redner, Walker 1984], Gibbs sampling [Asuncion et. al. 2011]
  - Local tests [Bresler et. al. 2008, Anadkumar et. al. 2012,]
  - Convex relaxations (e.g. Lasso) [Meinshausen, Bühlmann 2006, Ravikumar, Wainwright 2010]

Learning latent models

Goal: Given the observed data, learn structure and parameters of model.

Challenges:

- Identifiablity Many models can explain the observed data!
  - ICA: no edge between hidden nodes
  - LDA: hidden variables are drawn from a Dirichlet distribution
  - latent trees, graphical models with long cycles. [Anandkumar et.al. 2011, Choi et. al. 2011, Daskalakis et. al. 2006]
- Tractable learning algorithms:
  - Maximum likelihood (tractable on trees, NP-hard in general)
  - Expectation maximization [Redner, Walker 1984], Gibbs sampling [Asuncion et. al. 2011]
  - Local tests [Bresler et. al. 2008, Anadkumar et. al. 2012,]
  - Convex relaxations (e.g. Lasso) [Meinshausen, Bühlmann 2006, Ravikumar, Wainwright 2010]

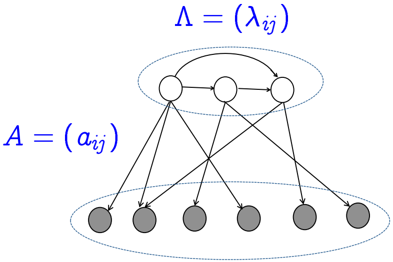
Learning latent models

Goal: Given the observed data, learn structure and parameters of model.

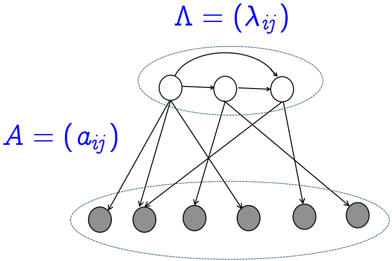
Challenges:

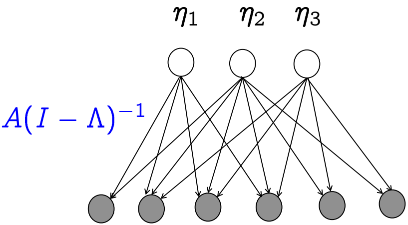
- Identifiablity Many models can explain the observed data!
  - ICA: no edge between hidden nodes
  - LDA: hidden variables are drawn from a Dirichlet distribution
  - latent trees, graphical models with long cycles. [Anandkumar et.al. 2011, Choi et. al. 2011, Daskalakis et. al. 2006]
- Tractable learning algorithms:
  - Maximum likelihood (tractable on trees, NP-hard in general)
  - Expectation maximization [Redner, Walker 1984], Gibbs sampling [Asuncion et. al. 2011]
  - Local tests [Bresler et. al. 2008, Anadkumar et. al. 2012,]
  - Convex relaxations (e.g. Lasso) [Meinshausen, Bühlmann 2006, Ravikumar, Wainwright 2010]

An example



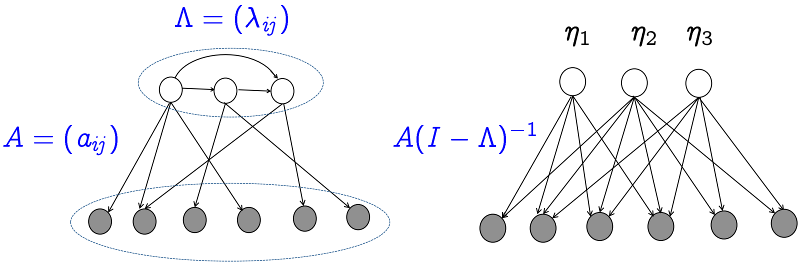
An example





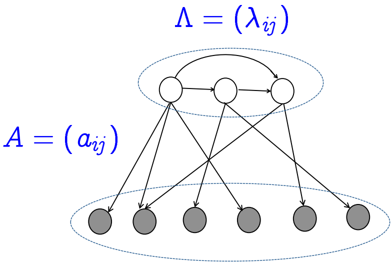
$$\begin{cases} x &= Ah + \epsilon \\ h &= \Lambda h + \eta \end{cases} \Longrightarrow x = A(I - \Lambda)^{-1}\eta + \epsilon$$

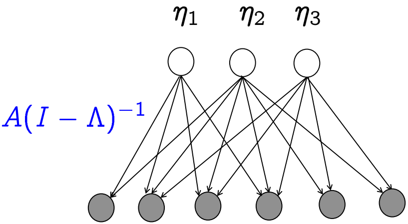
An example



A prudent restriction on the model

An example





A prudent restriction on the model

broadly applicable

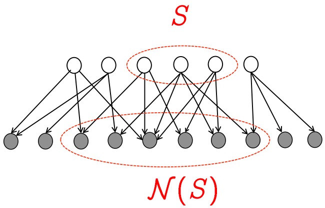
tractable learning methods

Sufficient conditions for identifiability

Task: Recover A

Structural Condition: (Additive) Graph Expansion

$$\boxed{|\mathcal{N}(S)| \geq |S| + d_{\max}, \text{ for all } S \subset \mathcal{H}}$$



Parametric Condition: Generic Parameters

$$\boxed{\|Av\|_0 > |\mathcal{N}_A(\mathrm{supp}(v))| - |\mathrm{supp}(v)|}$$

Identifiability result

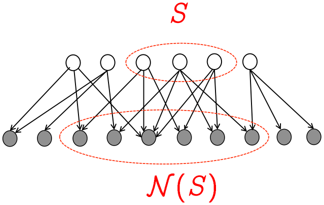
Under above conditions, A can be uniquely recovered from $\mathbb{E}[xx^\top]$.

Sufficient conditions for identifiability

Task: Recover A

Structural Condition: (Additive) Graph Expansion

$$\boxed{|\mathcal{N}(S)| \geq |S| + d_{\max}, \text{ for all } S \subset \mathcal{H}}$$



Parametric Condition: Generic Parameters

$$\boxed{\|Av\|_0 > |\mathcal{N}_A(\mathrm{supp}(v))| - |\mathrm{supp}(v)|}$$

Identifiability result

Under above conditions, A can be uniquely recovered from $\mathbb{E}[xx^\top]$.

Intuition

- Denoising the moment: $\mathbb{E}[xx^\mathsf{T}] = A\mathbb{E}[hh^\mathsf{T}]A^\mathsf{T} + \mathbb{E}[\epsilon\epsilon^\mathsf{T}]$

Intuition

- Denoising the moment: $\mathbb{E}[xx^{\mathsf{T}}] = \underbrace{A\mathbb{E}[hh^{\mathsf{T}}]A^{\mathsf{T}}}_{lowrank} + \underbrace{\mathbb{E}[\epsilon\epsilon^{\mathsf{T}}]}_{diagonal}$

Intuition

- Denoising the moment: $A\mathbb{E}[hh^{\mathsf{T}}]A^{\mathsf{T}}$

Intuition

- Denoising the moment: $A\mathbb{E}[hh^{\mathsf{T}}]A^{\mathsf{T}}$

- For non-degenerate $\mathbb{E}[hh^{\mathsf{T}}]$, we know Col($A$).

Intuition

- Denoising the moment: $A\mathbb{E}[hh^{\mathsf{T}}]A^{\mathsf{T}}$
- For non-degenerate $\mathbb{E}[hh^{\mathsf{T}}]$, we know Col($A$).
- Under above conditions, sparsest vectors in Col(A) are columns of A.

Intuition

- Denoising the moment: $A\mathbb{E}[hh^{\mathsf{T}}]A^{\mathsf{T}}$
- For non-degenerate $\mathbb{E}[hh^{\mathsf{T}}]$, we know Col($A$).
- Under above conditions, sparsest vectors in Col(A) are columns of A.

[Spielman, Wang, Wright 2012]

Intuition

- Denoising the moment: $A\mathbb{E}[hh^\mathsf{T}]A^\mathsf{T}$

- For non-degenerate $\mathbb{E}[hh^\mathsf{T}]$, we know Col($A$).

- Under above conditions, sparsest vectors in Col(A) are columns of A.

Exhaustive search

1. Let $U = \text{Col}(A\mathbb{E}[hh^\mathsf{T}]A^\mathsf{T})$
2. $\min_{z \neq 0} \|Uz\|_0$

A tractable algorithm

Task: Recover A from $U = \mathrm{Col}(A\mathbb{E}[hh^\top]A^\mathsf{T})$.

TWMLearn

1. Let $U = \mathrm{Col}(A\mathbb{E}[hh^\mathsf{T}])A^\mathsf{T} \in \mathbb{R}^{n\times k}$.
2. Solve
$$\min_z \|Uz\|_1, \quad (e_i^\mathsf{T} U)z = 1\,.$$
3. Set $s_i = Uz$, and $\mathcal{S} = \{s_1, \dots, s_n\}$.
4. Return a maximal full rank subset of $\mathcal{S}$.

Under "reasonable" conditions, the above program exactly recovers A

Learning latent space parameters

Recall so far ...

Recovered A

- from second order moment $\mathbb{E}[xx^\mathsf{T}]$
- under no assumption on the hidden variables!

What hidden structures can be learnt from low order observed moments?

Learning latent space parameters

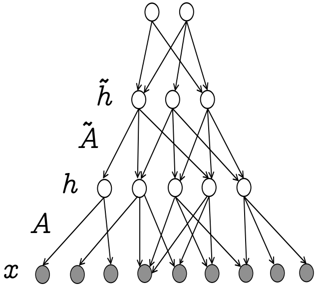
Recall so far ...

Recovered A

- from second order moment $\mathbb{E}[xx^\mathsf{T}]$
- under no assumption on the hidden variables!

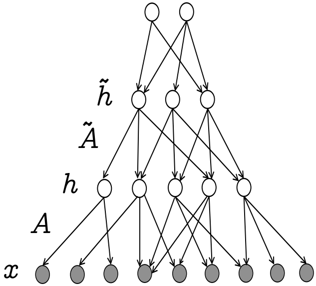
What hidden structures can be learnt from low order observed moments?

Multi-level DAGs



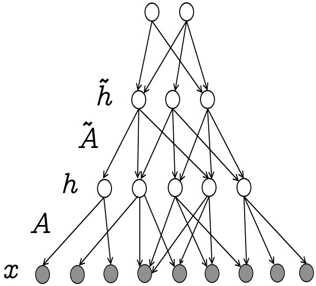
$$\mathbb{E}[xx^{\mathsf{T}}] = A\mathbb{E}[hh^{\mathsf{T}}]A^{\mathsf{T}} + \mathbb{E}[\epsilon\epsilon^{\mathsf{T}}]$$

Multi-level DAGs



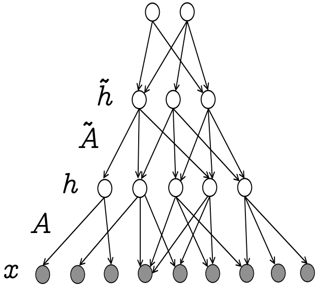
$$A\mathbb{E}[hh^{\mathsf{T}}]A^{\mathsf{T}}$$

Multi-level DAGs



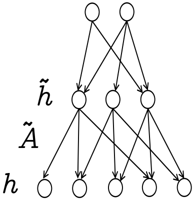
$$\color{red}{A}\color{black}{\mathbb{E}[hh^\mathsf{T}]}\color{red}{A^\mathsf{T}}$$

Multi-level DAGs



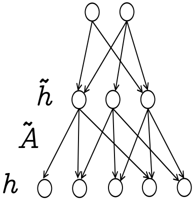
$$A^\dagger A \mathbb{E}[hh^\mathsf{T}] A^\mathsf{T} (A^\dagger)^\mathsf{T}$$

Multi-level DAGs



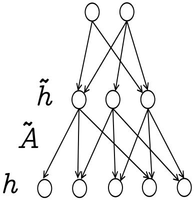
$$\mathbb{E}[hh^\mathsf{T}] = \tilde{A}\mathbb{E}[\tilde{h}\tilde{h}^\mathsf{T}]\tilde{A}^\mathsf{T} + \mathbb{E}[\tilde{\epsilon}\tilde{\epsilon}^\mathsf{T}]$$

Multi-level DAGs



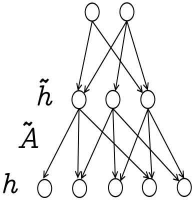
$$\tilde{A}\mathbb{E}[\tilde{h}\tilde{h}^{\mathsf{T}}]\tilde{A}^{\mathsf{T}}$$

Multi-level DAGs



$$\tilde{A}\mathbb{E}[\tilde{h}\tilde{h}^\mathsf{T}]\tilde{A}^\mathsf{T}$$

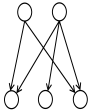
Multi-level DAGs



$$\tilde{A}^\dagger \tilde{A}\mathbb{E}[\tilde{h}\tilde{h}^\mathsf{T}]\tilde{A}^\mathsf{T}(\tilde{A}^\dagger)^\mathsf{T}$$

Multi-level DAGs

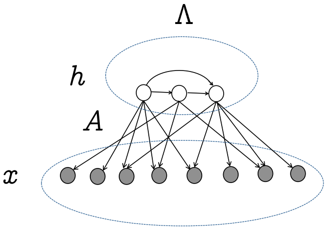
$\tilde{h}$



$$\mathbb{E}[\tilde{h}\tilde{h}^{\mathsf{T}}]$$

Linear structural equations

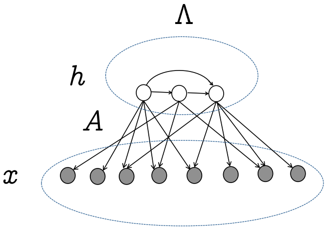
- Recall $\boxed{x = Ah + \epsilon}$
- Now additionally A is full rank (each hidden nodes has at least one observed neighbor)
- Linear dependence among hidden node: $h_j = \sum_{i \in \mathsf{PA}_j} \lambda_{ji} h_i + \eta_j$ (in matrix form $\boxed{h = \Lambda h + \eta}$)
- Noise variables η_j are uncorrelated.



Spectral approach for learning

Linear structural equations

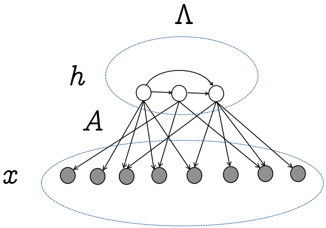
- Recall $\boxed{x = Ah + \epsilon}$
- Now additionally A is full rank (each hidden nodes has at least one observed neighbor)
- Linear dependence among hidden node: $h_j = \sum_{i \in \mathsf{PA}_j} \lambda_{ji} h_i + \eta_j$ (in matrix form $\boxed{h = \Lambda h + \eta}$)
- Noise variables η_j are uncorrelated.



Spectral approach for learning

Linear structural equations

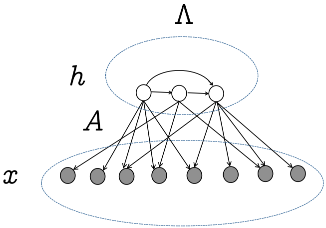
- Recall $\boxed{x = Ah + \epsilon}$
- Now additionally A is full rank (each hidden nodes has at least one observed neighbor)
- Linear dependence among hidden node: $h_j = \sum_{i \in \mathsf{PA}_j} \lambda_{ji} h_i + \eta_j$ (in matrix form $\boxed{h = \Lambda h + \eta}$)
- Noise variables η_j are uncorrelated.



Spectral approach for learning

Linear structural equations

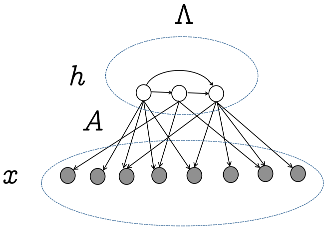
- Recall $\boxed{x = Ah + \epsilon}$
- Now additionally A is full rank (each hidden nodes has at least one observed neighbor)
- Linear dependence among hidden node: $h_j = \sum_{i \in \mathsf{PA}_j} \lambda_{ji} h_i + \eta_j$ (in matrix form $\boxed{h = \Lambda h + \eta}$)
- Noise variables η_j are uncorrelated.



Spectral approach for learning

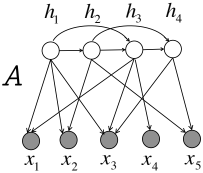
Linear structural equations

- Recall $\boxed{x = Ah + \epsilon}$
- Now additionally A is full rank (each hidden nodes has at least one observed neighbor)
- Linear dependence among hidden node: $h_j = \sum_{i \in \mathsf{PA}_j} \lambda_{ji} h_i + \eta_j$ (in matrix form $\boxed{h = \Lambda h + \eta}$)
- Noise variables η_j are uncorrelated.



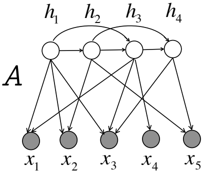
Spectral approach for learning

Learning Λ: idea

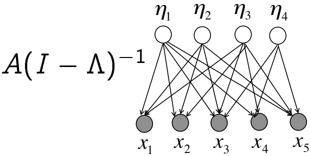


$$x = Ah + \epsilon$$
$$h = \Lambda h + \eta$$

Learning Λ: idea



$$x = Ah + \epsilon$$
$$h = \Lambda h + \eta$$



$$x = A(I-\Lambda)^{-1}\eta + \epsilon$$

Learning Λ: idea

- Employ spectral approach to learn $A(I-\Lambda)^{-1}$
  - second order moment:

$$\boxed{\mathbb{E}[xx^{\mathsf{T}}] = A(I-\Lambda)^{-1}\mathbb{E}[\eta\eta^{\mathsf{T}}](A(I-\Lambda))^{\mathsf{T}} + \mathbb{E}[\epsilon\epsilon^{\mathsf{T}}]}$$

  - third order moment:

$$\boxed{\mathbb{E}[xx^{\mathsf{T}}\langle\xi, x\rangle] = A(I-\Lambda)^{-1}\mathbb{E}[\eta\eta^{\mathsf{T}}\langle\eta, A^{\mathsf{T}}\xi\rangle](A(I-\Lambda))^{\mathsf{T}} + \mathbb{E}[\epsilon\epsilon^{\mathsf{T}}\langle\xi, \epsilon\rangle]}$$

Learning Λ: idea

- Employ spectral approach to learn $A(I-\Lambda)^{-1}$
  - second order moment:

$$\boxed{\mathbb{E}[xx^\mathsf{T}] = A(I-\Lambda)^{-1}\mathbb{E}[\eta\eta^\mathsf{T}](A(I-\Lambda))^\mathsf{T} + \mathbb{E}[\epsilon\epsilon^\mathsf{T}]}$$

  - third order moment:

$$\boxed{\mathbb{E}[xx^\mathsf{T}\langle\xi,x\rangle] = A(I-\Lambda)^{-1}\mathbb{E}[\eta\eta^\mathsf{T}\langle\eta,A^\mathsf{T}\xi\rangle](A(I-\Lambda))^\mathsf{T} + \mathbb{E}[\epsilon\epsilon^\mathsf{T}\langle\xi,\epsilon\rangle]}$$

- Simultaneous diagonalization of the moments
  (through SVD or tensor decompositions)

[Anandkumar, Foster, Hsu, Kakade, Liu 2012]

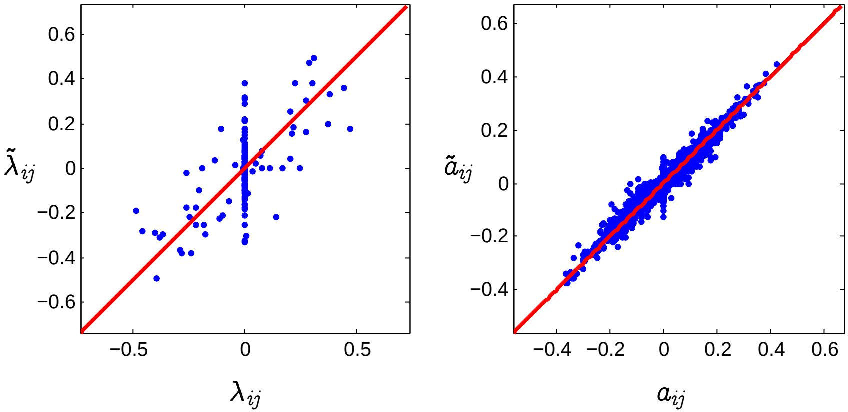
[Anandkumar, Ge, Hsu, Kakade 2012]

- "Col(A) = Col($A(I-\Lambda)^{-1}$)" + "expansion property" $\Rightarrow$ A and Λ

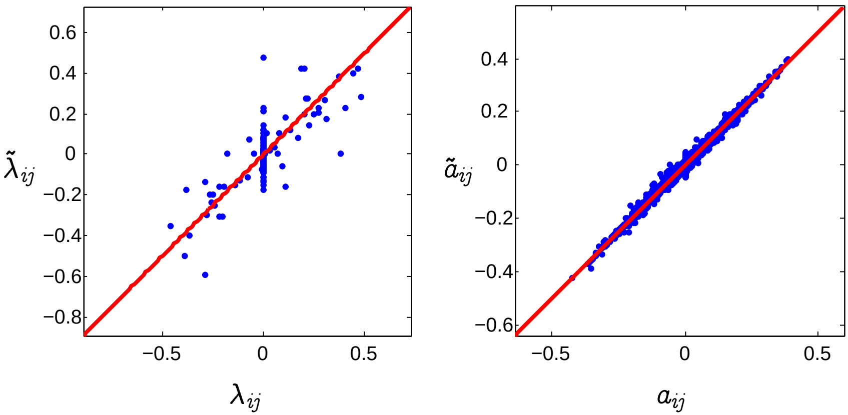
Experiment

- $k = 25$ hidden nodes and $n = 150$ observed nodes
- Bernoulli-Gaussian model ($p = 0.3$), total number of edges $= 1177$.
- Noise variables distributed as exponential, poisson, chi-2, Gaussian with mean zero and variances chosen randomly in $[0.5, 1]$.

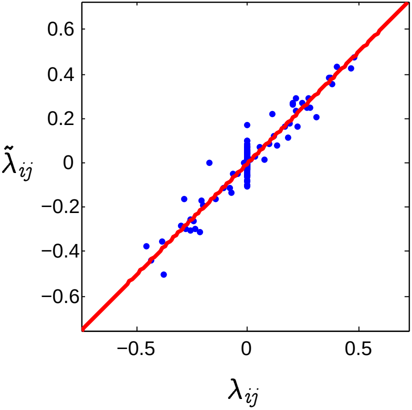
number of samples = 25,000

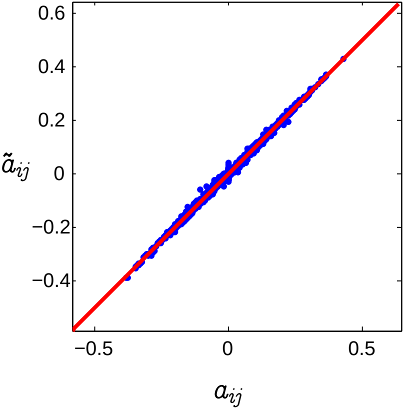


number of samples = 100,000



number of samples = 400,000





Conclusion

- Considered learning latent models with arbitrary hidden variable dependencies.
- Constraint on the model: expansion of bipartite graph from hidden to observed layer, generic parameter and non-degeneracy.
- Established identifiability of A under no assumption but non-degeneracy of the hidden variables!
- Recovering A through ℓ_1 optimization.
- Can be used to learn topic-word matrix under the expansion constraint and arbitrary topic dependencies.
- Learning the hidden space parameters and structure for multi-level DAGs and linear structural equations.

You are welcome to visit our poster presentation (Paper ID: 146)!

Thanks!