Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures

VINCENT Y. F. TAN, ANIMASHREE ANANDKUMAR AND ALAN S. WILLSKY[†]

Stochastic Systems Group, Laboratory for Information and Decision Systems

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

Cambridge, MA 02139, USA

[†]{vtan, animakum, willsky}@mit.edu

Introduction

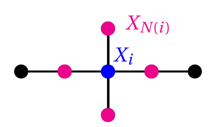
- Learning the structure and interdependencies among a large collection of variables is an important generic task
- Challenging when the dimensionality of the data is large compared to the number of samples
- Need to find the right balance between data fidelity and overfitting to the model
- The work focuses on learning tree-structured Gaussian graphical models, which have a fixed number of parameters
- We derive the error exponent for learning the tree structure
- How do the structure $\mathcal{E}_p$ and the parameters of the original model affect the error exponent K_p?
- What are the extremal tree distributions that maximize and minimize the exponent?

Preliminaries

Graphical Models

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random vector, where each variable X_i corresponds to node $i \in \mathcal{V}$ in $\mathcal{G}$. We say that $\mathbf{X}$ is Markov on $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if for every $i \in \mathcal{V}$,

$$X_i \perp X_{\mathcal{V}\setminus(N(i)\cup\{i\})} \mid X_{N(i)}$$



The above is known as the local Markov property.

Tree-Structured Graphical Models

If $\mathcal{G}$ is a tree, then the joint distribution of $\mathbf{X}$ factorizes as

$$p(x_1, \ldots, x_d) = \prod_{i\in\mathcal{V}} p(x_i) \prod_{(i,j)\in\mathcal{E}} \frac{p(x_i, x_j)}{p(x_i)p(x_j)}.$$

This is a generalization of Markov chains: If $X_1 - X_2 - X_3$ form a Markov chain in that order, then $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$.

Gaussian Graphical Models

In this work, we focus on Gaussian graphical models (GMRFs), i.e.,

$$p(x_1, \ldots, x_d) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}^{-1}\mathbf{x}\right).$$

$[\mathbf{\Sigma}^{-1}]_{i,j} = 0$ iff $(i,j) \notin \mathcal{E}_p$.

The Chow-Liu Algorithm

We are given samples $\mathbf{x}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ drawn i.i.d. from p, Markov on $\mathcal{T}_p = (\mathcal{V}, \mathcal{E}_p)$. Solve the following reverse I-projection problem:

$$\widehat{\mathcal{E}}(\mathbf{x}^n) := \operatorname*{argmin}_{q\in\text{Trees}} D(\widehat{p}\,||\,q)$$

where $\widehat{p} = \mathcal{N}(\mathbf{x}; \mathbf{0}, \widehat{\mathbf{\Sigma}})$ and $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^T$. Chow and Liu (1968) showed that

$$\widehat{\mathcal{E}}(\mathbf{x}^n) = \operatorname*{argmax}_{\mathcal{E}\in\text{Trees}} \sum_{(i,j)\in\mathcal{E}} I(\widehat{p}_{i,j}),$$

where the edge weights are the empirical MI $I(\widehat{p}_{i,j}) = -\frac{1}{2}\log(1-\widehat{\rho}_{i,j}^2)$. Can be solved via a max-weight spanning tree procedure.

Problem Statement

Define the error event as

$$\mathcal{A}_n := \{\mathbf{x}^n : \widehat{\mathcal{E}}(\mathbf{x}^n) \neq \mathcal{E}_p\}$$

Computing and analyze the error exponent:

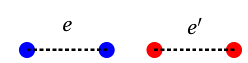
$$K_p := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\mathcal{A}_n).$$

Quantifies the relative ease in learning the model

Remarks

- Exhaustive search for the closest tree to p is not feasible: d^{d-2} trees with d nodes
- If ranking of empirical MI is correct, then $\{\widehat{\mathcal{E}}(\mathbf{x}^n) = \mathcal{E}_p\}$

Analyzing Crossover Events



Imagine that there are two pairs of nodes $e, e' \in \binom{\mathcal{V}}{2}$ such that

$$I(p_e) > I(p_{e'}).$$

Consider the crossover event of the empirical MI

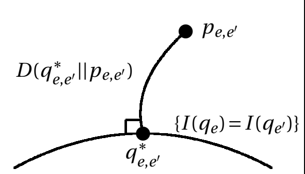
$$\{I(\widehat{p}_e) \leq I(\widehat{p}_{e'})\}.$$

Definition: Crossover Rate

$$J_{e,e'} := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left(I(\widehat{p}_e) \leq I(\widehat{p}_{e'})\right).$$

Lemma: The crossover rate is

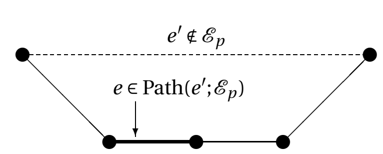
$$J_{e,e'} = \inf_{q\in\text{Gaussian}: I(q_e)=I(q_{e'})} D(q\,||\,p_{e,e'})$$



- Proof by Sanov's theorem (Large deviations)
- Non-convex optimization ☹

Error Exponent for Structure Learning

- Need to consider only one dominant crossover event by large deviations theory
- Identify the crossover event with the minimum rate $J_{e,e'}$
- Tree constraint must be satisfied



Theorem: The error exponent for learning tree-structured Gaussian graphical models is

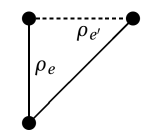
$$K_p = \min_{e'\notin\mathcal{E}_p} \min_{e\in\text{Path}(e';\mathcal{E}_p)} J_{e,e'}$$

Euclidean Approximations

- Optimization for crossover rate $J_{e,e'}$ is non-convex
- Consider parameters of Gaussians that are hard for learning
- Lend more insight into how errors occur in structure learning

Definition: The joint distribution $p_{e,e'} = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Sigma}_{e,e'})$ is ϵ-very noisy if

$$-\epsilon < |\rho_e| - |\rho_{e'}| < \epsilon,$$
$$|\rho_e| \approx |\rho_{e'}|$$

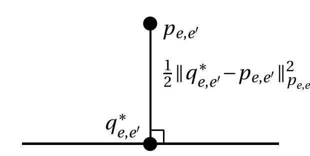


Theorem: When ϵ is small, the crossover rate can be approximated as

$$\widetilde{J}_{e,e'} = \frac{(I(p_e) - I(p_{e'}))^2}{2\text{Var}(S_e - S_{e'})}$$

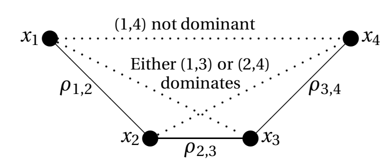
where

$$S_{i,j}(X_i, X_j) := \log\frac{p(X_i, X_j)}{p(X_i)p(X_j)}$$



More intuitive expression for the crossover rate ☺

Simplification of Error Exponent



Note by Markovianity that

$$\rho_e = \prod_{e'\in\text{Path}(e;\mathcal{E}_p)} \rho_{e'}, \qquad \Rightarrow \qquad \rho_{1,4} = \rho_{1,2}\rho_{2,3}\rho_{3,4}$$

Lemma: Data-processing inequality for crossover rates:

$$\widetilde{J}(\rho_{1,2}, \rho_{1,3}) \leq \widetilde{J}(\rho_{1,2}, \rho_{1,4}), \qquad \forall |\rho_{1,3}| \geq |\rho_{1,4}|$$

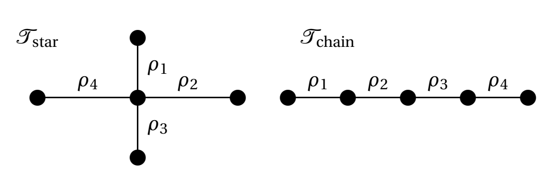
As a result, the error exponent $\widetilde{K}_p$ can be expressed as:

$$\widetilde{K}_p = \min_{e\in\mathcal{E}_p} \widetilde{J}(\rho_e, \rho_e\rho_e^*), \qquad \rho_e^* := \max\{|\rho_{\tilde{e}}| : \tilde{e}\in\mathcal{E}_p, \tilde{e}\sim e\}$$

Only $O(d)$ computations required! ☺

Extremal Structures

- Let $\boldsymbol{\rho} := [\rho_1, \ldots, \rho_{d-1}]$ be a fixed vector of correlation coefficients
- Uniquely determines parameters of a Gaussian graphical model



- Find the extremal distributions

$$p_{\max,\boldsymbol{\rho}} := \operatorname*{argmax}_{q \text{ has cc } \boldsymbol{\rho} \text{ on edges}} \widetilde{K}_q \qquad p_{\min,\boldsymbol{\rho}} := \operatorname*{argmin}_{q \text{ has cc } \boldsymbol{\rho} \text{ on edges}} \widetilde{K}_q$$

Main Result

Instead of characterizing the extremal distributions $p_{\min,\boldsymbol{\rho}}$ and $p_{\max,\boldsymbol{\rho}}$, characterize the structures that maximize and minimize the approximate error exponent $\widetilde{K}_p$. For fixed $\boldsymbol{\rho}$ on edges:

Theorem: The tree structure that minimizes the error exponent

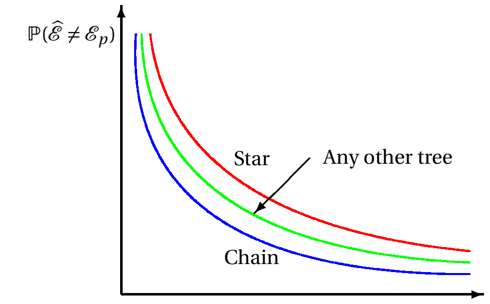
$$\mathcal{T}_{p_{\min,\boldsymbol{\rho}}} = \mathcal{T}_{\text{star}}$$

If, in addition, $|\rho_{e_i}| \leq 0.63$ for all edges $i = 1, \ldots, d-1$, then

$$\mathcal{T}_{p_{\max,\boldsymbol{\rho}}} = \mathcal{T}_{\text{chain}}$$

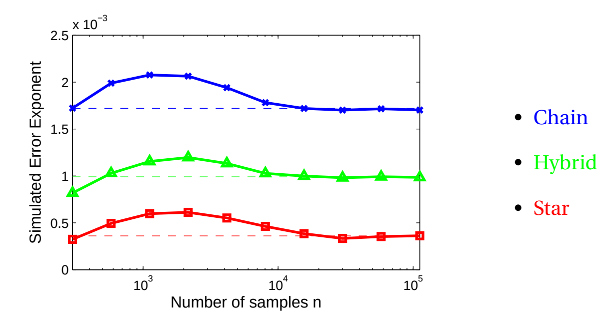
Remarks

- In the star, nodes are strongly correlated (no correlation decay)
- In the chain, there are many weakly correlated pairs of nodes
- Hardest to learn the star; Easiest to learn the chain
- Extremal structures independent of correlation coefficients
- Result means that in the limit of large n,



Experiments

- Learned Chow-Liu trees when the original structure with $d = 10$ nodes is a chain, star or hybrid graph
- Simulated the simulated error probability and error exponent.



References

- C. K. Chow and C. N. Liu. "Approximating discrete probability distributions with dependence trees". Trans. on IT, May 1968.
- V. Y. F. Tan, A. Anandkumar and A. S. Willsky. "Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures." Trans. on SP, May 2010.