On Learning Discrete Graphical Models using Group-Sparse Regularization

Ali Jalali[†], Pradeep Ravikumar[‡], Vishvas Vasuki[‡] and Sujay Sanghavi[†]

[†]*Electrical and Computer Engineering Department,* [‡]*Computer Sciences Department, The University of Texas at Austin*

Introduction

◇ Problem Setup

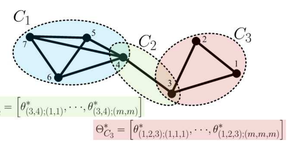
- Random variables $X = (X_1, X_2, \cdots X_p)$ taking values in $[m] = \{1, 2, \cdots, m\}$
- Joint PDF $\mathbb{P}(X) \propto \exp\left(\sum_{C\in\mathcal{C}} \phi_C(X_C)\right)$ for some $\mathcal{C} \subseteq 2^{[p]}$ and potentials $\phi_C(\cdot)$
- Potential Functions $\phi_C(X_C) = \sum_{v\in[m]^{|C|}} \theta^*_{C;v}\, \mathbb{I}[x_C = v]$

Given n iid samples $X^{(i)}$, learn Θ^*_C for all $C \in 2^{[p]}$

◇ Graphical Model

- Define the graph $\mathcal{G}$ with p vertices corresponding to variables
- If $\Theta^*_C \neq 0$ then make a complete graph on vertices X_C

Example:
$\mathbb{P}(X) \propto f_1(X_{C_1}; \Theta^*_{C_1}) f_2(X_{C_2}; \Theta^*_{C_2}) f_3(X_{C_3}; \Theta^*_{C_3})$

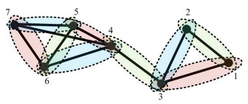


Given n ($\propto \log(p)$) iid samples $X^{(i)}$, learn the *sparse* graph $\mathcal{G}$

◇ Pairwise Graphical Model

- Only cliques of size 2 exist
- Assign $\Theta_{rt} \in \mathbb{R}^{m^2}$ to the edge (r,t) intersecting X_r and X_t

Example:
$\mathbb{P}(X) \propto \prod_{i=1}^{10} f_i(X_{C_i}; \Theta^*_{C_i})$



Methodology

◇ Learning Neighborhoods

- For each vertex learn its neighborhood consistently
- Combine all neighborhoods to get the graph

◇ Conditional Likelihood

- Fixing a node X_r;

$$\mathbb{P}\left[X_r = j | X_{\backslash r}\right] = \frac{\exp\left(\sum_{t\in[p]-\{r\}} \sum_{k=1}^{m} \theta^*_{rt;jk}\, \mathbb{I}[X_t = k]\right)}{\sum_{j=1}^{m} \exp\left(\sum_{t\in[p]-\{r\}} \sum_{k=1}^{m} \theta^*_{rt;jk}\, \mathbb{I}[X_t = k]\right)}$$

- Likelihood function

$$\mathcal{L}(\Theta_r) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\mathbb{P}\left[X_r = x_r^{(i)} | X_{\backslash r} = x_{\backslash r}^{(i)}\right]\right)$$

Our Algorithm

Algorithm 1 (Discrete Graphical Model Learner)**.** *For a fixed vertex r, find $\widehat{\Theta}_r$; the solution of the following convex optimization program*

$$\textit{Minimize:}\quad \mathcal{L}(\Theta_r) + \lambda \sum_{t\in[p]-\{r\}} \|\Theta_{rt}\|_2$$

Output the neighborhood $\widehat{\mathcal{N}}_r = \left\{t : \left\|\widehat{\Theta}_{rt}\right\|_0 \neq 0\right\}$.

Pairwise Case

◇ Assumptions

Let $Q^*_r = \mathbb{E}\left[\nabla^2 \log(\mathbb{P}[X_r | X_{\backslash r}])\right]$ be the Fisher information matrix.

- (A1) Minimum curvature: $\Lambda_{\min}(Q^*_r) \geq C_{\min} > 0$.
- (A2) Incoherence: $\left\|Q^*_{\mathcal{N}_r^c\mathcal{N}_r}\left(Q^*_{\mathcal{N}_r\mathcal{N}_r}\right)^{-1}\right\|_{\infty,2} \leq \frac{1-2\alpha}{\sqrt{|\mathcal{N}_r|}}$ for some $\alpha \in (0, 1/2]$.

Theorem 1 (Pairwise model sparsistency)**.** *Under assumptions (A1)-(A2), suppose*

$$n \geq K m^2 |\mathcal{N}_r|^2 \log(m^2 p)$$

and $\lambda \geq \frac{8(2-\alpha)}{\alpha}\left(\sqrt{\frac{\log(p)}{n}} + \frac{m-1}{4\sqrt{n}}\right)$. *Then, Algorithm 1, with probability* $1 - c_1\exp(-c_2\lambda^2 n)$, *excludes all false neighbors and includes all neighbors t provided that* $\|\Theta_{rt}\|_2 \geq \frac{10}{C_{\min}}\lambda$.

General Higher-Order Case

◇ Further Assumption

- (A3) Bounded Mistmatch: Non-Pairwise (highr order) parameters $\Theta^*_{P^c}$ satisfy $\left\|\Theta^*_{P^c}\right\|_1 \leq K\frac{C^2_{\min}}{(m-1)|\mathcal{N}_r|}$.

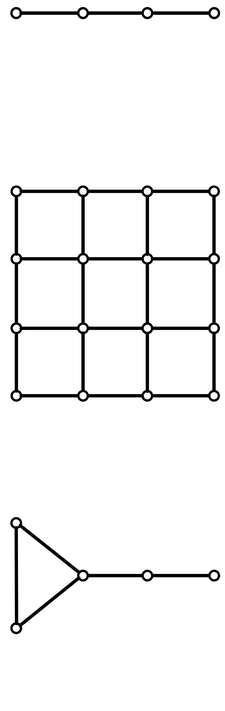
Theorem 2 (Higher-Order model sparsistency)**.** *Under assumptions (A1)-(A3), suppose*

$$n \geq K m^2 |\mathcal{N}_r|^{1.5c+0.5} \log(m^c p^{c-1})$$

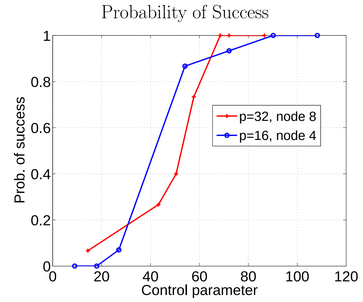
and $\lambda \geq \frac{8(2-\alpha)}{\alpha}\left(\sqrt{\frac{\log(p)}{n}} + \frac{m-1}{4\sqrt{n}} + \frac{1}{4}\left\|\Theta^*_{P^c}\right\|_1\right)$. *Then, Algorithm 1, with probability* $1 - c_1\exp(-c_2(\lambda^2 - 2\left\|\Theta^*_{P^c}\right\|_1)n)$, *excludes all false neighbors and includes all neighbors t provided that* $\|\Theta_{rt}\|_2 \geq \frac{10}{C_{\min}}\lambda$.

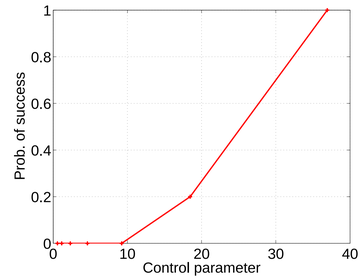
Numerical Experiments

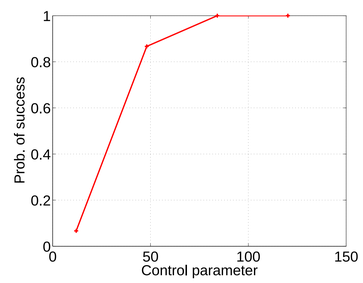
True Graphical Model



Probability of Success







Group-sparse regularization can efficiently learn general discrete graphical models.

References

[1] P. Ravikumar, M. J. Wainwright, and J. Lafferty. "High-dimensional ising model selection using ℓ_1-regularized logistic regression", Annals of Statistics, 38(3), 1287:1319.

[2] F. Bach, "Consistency of the group LASSO and multiple kernel learning", Journal of Machine Learning Research, 9:1179-1225, 2008.

[3] N. Meinshausen and P. Buhlmann. "High dimensional graphs and variable selection with the lasso", Annals of Statistics, 34(3), 2006.