# Learning Graphs with a Few Hubs


Rashish Tandon, Pradeep Ravikumar

Dept. of Computer Science, The University of Texas at Austin

THE UNIVERSITY OF TEXAS AT AUSTIN


## Overview

We consider the problem of **learning graphical models** (in particular, *ising* models) when the underlying graph can have nodes with a high degree; typically, high degree neighborhoods are hard to learn. Instead, we propose a quantitative criterion called the *sufficiency measure* that:

- Indicates inability to learn neighborhoods
- Can be estimated
- Allows learning graph from *recoverable* neighborhoods

## Introduction and Problem Setup

A **graphical model** with graph $G = (V, E)$ (with $|V| = p$) represents a **multivariate distribution**

- Nodes $i \in V$ correspond to **random variables** $X_i$ $(\in \mathcal{X})$
- Edges $E$ encode **markov independence** relationships

For the special case of a **pairwise ising model** with graph $G = (V, E)$:

$$\mathbb{P}_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{rt \in E} \theta_{rt} x_r x_t \right\} \qquad (1)$$

- $x \in \{-1, 1\}^p$: Binary random variables
- $\theta_{rt} \in \mathbb{R}$: Weight on edge $(r, t) \in E$
- $\theta \in \mathbb{R}^{\binom{p}{2}}$: Set of edge weights
- $\mathbb{P}_\theta(x)$: The probability mass function (pmf)
- $Z(\theta)$: The normalization factor

**Applications**: Statistical Physics, Natural Language Processing, Image Analysis, Spatial Statistics etc.

The **Structure Learning** problem: Recover underlying **graph structure**, given data

- Observe $n$ samples $D = \{x^{(1)}, \dots, x^{(n)}\}$ drawn i.i.d from $\mathbb{P}_{\theta^*}$, with graph $G^* = (V, E^*)$
- Relation between $\theta^*$ and $E^*$ : $E^* = \{(r,t) \in V \times V \mid \theta^*_{rt} \neq 0\}$
- Goal is to provide an estimate, $\widehat{E}_n$, of $E^*$
- High-dimensional regime: $p >> n$
- Estimator is **sparsistent** if $\mathbb{P}\left[\widehat{E}_n = E^*\right] \to 1$ as $n \to \infty$

## $\ell_1$-estimator: Overview[1]

Estimates the **neighbourhood** for **each node**. Combines neighborhoods using an AND/OR rule.

- For each node $r \in V$, let $\theta_{\setminus r} = \{\theta_{rt} \mid t \in V, t \neq r\}$
- True neighbourhood : $\mathcal{N}^*(r) = \text{Support}\left(\theta^*_{\setminus r}\right)$

## $\ell_1$-estimator: Overview[1]

Minimize **negative conditional log-likelihood** (logistic likelihood) for each node $r \in V$ subject to $\ell_1$-regularization *i.e.*

$$\widehat{\theta}_{\setminus r}(\lambda; D) = \underset{\theta_{\setminus r} \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \underbrace{\mathcal{L}(\theta_{\setminus r}; D)}_{\text{logistic likelihood}} + \lambda \underbrace{\|\theta_{\setminus r}\|_1}_{\ell_1 \text{ norm}} \right\}$$

**Neighborhood estimate** : $\widehat{\mathcal{N}}_\lambda(r; D) = \text{Support}\left(\widehat{\theta}_{\setminus r}(\lambda; D)\right)$

- Strong statistical guarantees under incoherence
- If $n = \Omega\left(d_r^3 \log p\right)$, then w.h.p. recovers the neighborhood accurately *i.e.* $\widehat{\mathcal{N}}_\lambda(r; D) = \mathcal{N}^*(r)$. $d_r$ = degree of node $r$.
- **For entire graph**: $n = \Omega\left(d_{\max}^3 \log p\right)$ samples suffice
- **Key Problem**: Estimating a large neighborhood requires many samples
- Consider a **star graph**: *one hub node* and *p spoke nodes*
  - Overall recovery: $\Omega(p^3 \log p)$ samples
  - Recovering neighborhood of spoke nodes: $\Omega(\log p)$ samples only !

## Sufficiency Measure

A **quantitative indicator** of *difficult* estimation.

- For $r \in V$, $t \in V \setminus r$, define : $p_{r,n,\lambda}(t) = \mathbb{P}\left(t \in \widehat{\mathcal{N}}_\lambda(r; D)\right)$
- $p_{r,n,\lambda}(t)$ = Probability of $t$ appearing in neighborhood estimate of $r$
- When we have sufficient samples, $p_{r,n,\lambda}(t) \approx 0$ or $p_{r,n,\lambda}(t) \approx 1$

### Sufficiency Measure

$$\mathcal{M}_{r,n,\lambda} = \max_{t \in V \setminus r} p_{r,n,\lambda}(t)\left(1 - p_{r,n,\lambda}(t)\right)$$

- When we have sufficient samples, $\mathcal{M}_{r,n,\lambda} \approx 0$

## Estimating the Sufficiency Measure

- Consider $N$ sub-samples $\{D_1, \dots, D_N\}$ of $D$, each of size $b$
- Estimate $p_{r,b,\lambda}(t)$ as :

$$\widehat{p}_{r,b,\lambda}(t) = \frac{1}{N} \sum_{i \in [N]} \mathbb{I}\left(t \in \widehat{\mathcal{N}}_{b,\lambda}(r; D_i)\right)$$

where $\mathbb{I}(\cdot)$ is a $0-1$ indicator

- Sufficiency Measure estimate :

$$\widehat{\mathcal{M}}_{r,b,\lambda}(D) = \max_{t \in V \setminus r} \widehat{p}_{r,b,\lambda}(t)\left(1 - \widehat{p}_{r,b,\lambda}(t)\right)$$

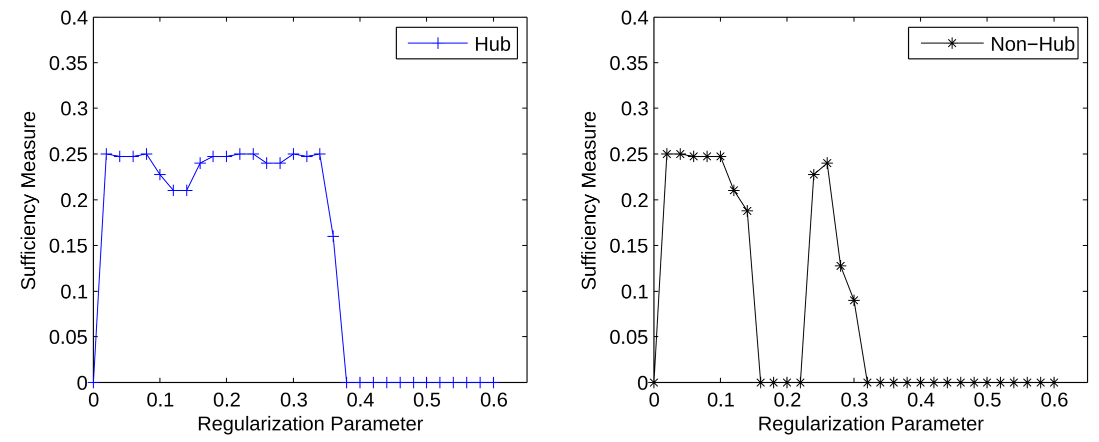
**Proposition:** For any $\delta \in (0, 1]$ and $\epsilon > 0$, if we have $n > \frac{18b}{\epsilon^2}\left[\log p + \log(4/\delta)\right]$ and $N \geq \lceil \frac{n}{b} \rceil$, then,

$$\mathbb{P}\left(|\widehat{\mathcal{M}}_{r,b,\lambda}(D) - \mathcal{M}_{r,b,\lambda}| \leq \epsilon\right) \geq 1 - \delta. \qquad (2)$$

## Properties of the Sufficiency Measure

**Definition [Non-Hub Node]:** Assume the edge weights $\theta^*$ satisfy incoherence. Then, any node $r \in V$ is called a **non-hub node** w.r.t. $n$ given samples if the $\ell_1$-estimator can recover its neighborhood exactly (*stated less formally*)

- Note that $n = \Omega\left(d_r^3 \log p\right)$ samples suffice for to recover neighborhood of node $r \in V$ (w.h.p.)
- Thus, **Non-hub nodes** (as defined above) $\sim$ **low degree nodes**
- Any node which is not a non-hub node is termed a **hub node**




(a) $\mathcal{M}_{r,b,\lambda}$ for hub node ($d_r = 19$) (b) $\mathcal{M}_{r,b,\lambda}$ for non-hub node ($d_r = 1$)

Figure 1: Behaviour of $\mathcal{M}_{r,b,\lambda}$ for non-hub nodes and hub-nodes in a star graph on $p = 100$ nodes

- Under some assumptions on $p_{r,b,\lambda}(t)$, the behaviour of $\mathcal{M}_{r,b,\lambda}$ can be characterized theoretically
- **Bell/Bump** : Between finite values $\lambda_l$ and $\lambda_u$, $\mathcal{M}_{r,b,\lambda}$ is always above a very small threshold
- **End of Bell/Bump** : At any point $\lambda$ where $\mathcal{M}_{r,b,\lambda}$ falls below this threshold, $\widehat{\mathcal{N}}_{b,\lambda}(r; D) \subseteq \mathcal{N}^*(r)$ w.h.p.
- Choosing the penultimate point where $\mathcal{M}_{r,b,\lambda}$ falls below a small threshold gives the best neighborhood recovery

## Using Sufficiency Measure for Estimation

- Input : Data $D := \{x^{(1)}, \dots, x^{(n)}\}$ , Regularization parameters $\Lambda := \{\lambda_1, \dots, \lambda_s\}$ , Sub-sample size $b$, No. of sub-samples $N$, Thresholds $t_l$ and $t_u$
- Output : $\widehat{E}$

**foreach** $r \in V$ **do**

  $\forall \lambda \in \Lambda$, Compute $\widehat{\mathcal{M}}_{r,b,\lambda}(D)$ and $\widehat{p}_{r,b,\lambda}(t; D)\ \forall t \in V \setminus r$

  $\lambda' \leftarrow$ Smallest $\lambda \in \Lambda$ s.t. $\widehat{\mathcal{M}}_{r,b,\lambda}(D) > t_u$

  $\Lambda \leftarrow \{\lambda \in \Lambda : \lambda > \lambda'\}$

  $\lambda_0 \leftarrow$ Smallest $\lambda \in \Lambda$ s.t. $\widehat{\mathcal{M}}_{r,b,\lambda}(D) < t_l$

  $\widehat{\mathcal{N}}(r) \leftarrow \left\{t \mid \widehat{p}_{r,b,\lambda_0}(t; D) \geq \frac{1+\sqrt{1-4t_l}}{2}\right\}$

$\widehat{E} \leftarrow \bigcup_{r \in V} \{(r,t) \mid t \in \widehat{\mathcal{N}}(r)\}$

## Guarantees

Suppose,

- $t_l = 2\exp(-c \log p)(1 - 2\exp(-c \log p)) + \epsilon$, $t_u = 1/4 - \epsilon$
- Sub-sample size $b = f(n) > c' d^3 \log p$
- Number of sub-samples $N \geq \lceil n/b \rceil$
- $n > 18b\left[\log p + \log(4/\delta)\right]/\epsilon^2$
- $|\theta^*_{st}| \geq c'' \sqrt{\frac{d \log p}{n}}\ \forall st \in E^*$

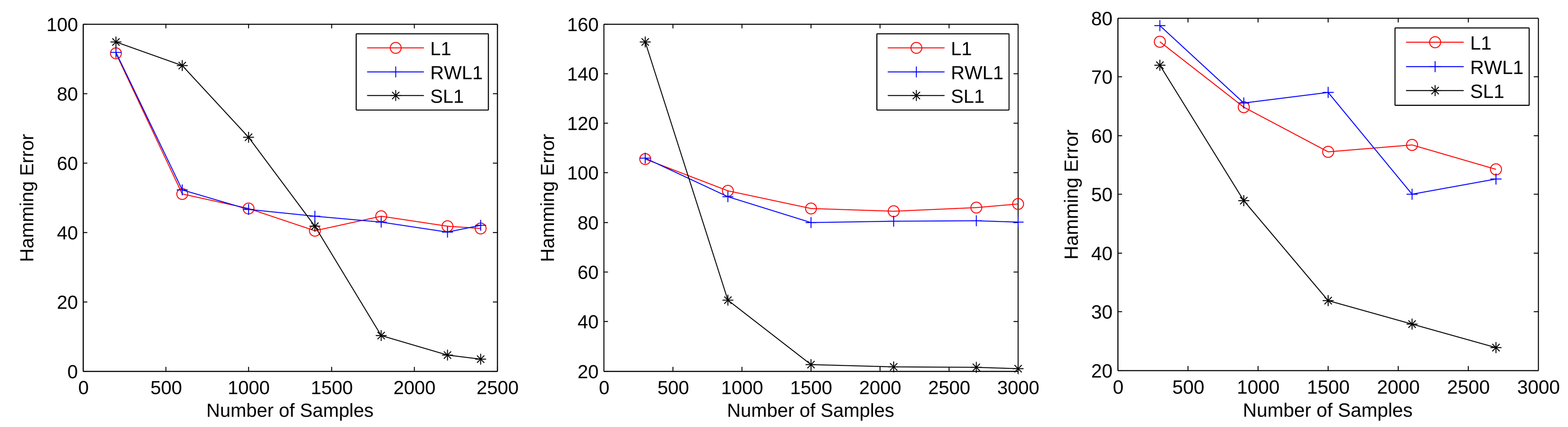
where $c, c', c''$ are constants

**Theorem:** Let $E_d$ be union of edges $(u,v)$ in $E^*$ s.t. either $d(u) \leq d$ or $d(v) \leq d$. Then,

$$\mathbb{P}\left(E_d \subseteq \widehat{E} \subseteq E^*\right) \geq 1 - 2\exp\left(-c''' \log p\right) - \delta. \qquad (3)$$

- Define **critical degree**, $d_c$ as the minimum degree $d'$ s.t. $\forall (s,t) \in E^*$, either $d(s) \leq d'$ or $d(t) \leq d'$
- To learn entire graph, in the above, pick $d = d_c$.
- **Example**: $d_c = 1$ for Star graph.

## Experiments




(a) Star Graph ($p = 100$, hub degree=19) (b) Hub+Grid Graph ($p = 83$, hub degree=12) (c) Preferential Attachment Graph ($p = 100$)

Figure 2: Plots of Average Hamming Error vs Number of Samples


**References**:

[1] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using $\ell_1$-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.