Learning network models from data: Practical methods and fundamental limits

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Based on joint works with:

John Lafferty (CMU)
Garvesh Raskutti (UC Berkeley)
Pradeep Ravikumar (UT Austin)
Prasad Santhanam (Univ. Hawaii)
Bin Yu (UC Berkeley)

Introduction

- Markov random fields (undirected graphical models): central to many applications in science and engineering:
    - communication, coding, and information theory
    - control theory, networking
    - machine learning and statistics
    - statistical signal processing
    - combinatorial optimization, theoretical computer science
    - computational biology

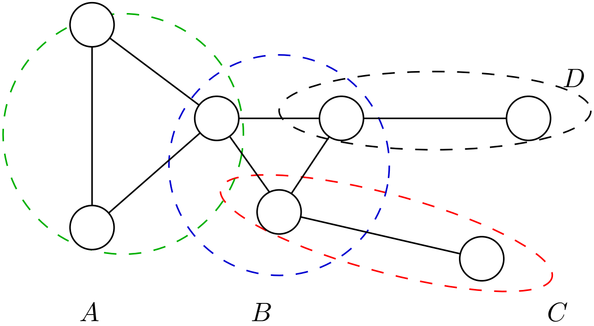
Introduction

- Markov random fields (undirected graphical models): central to many applications in science and engineering:
  - communication, coding, and information theory
  - control theory, networking
  - machine learning and statistics
  - statistical signal processing
  - combinatorial optimization, theoretical computer science
  - computational biology

- some core computational problems
  - *counting/integrating:* computing marginal distributions and data likelihoods
  - *optimization:* computing most probable configurations (or top M-configurations)
  - *model selection:* fitting and selecting models on the basis of data

What are graphical models?

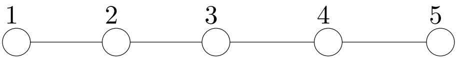
- Markov random field: random vector $(X_1, \ldots, X_p)$ with distribution factoring according to a graph $G = (V, E)$:

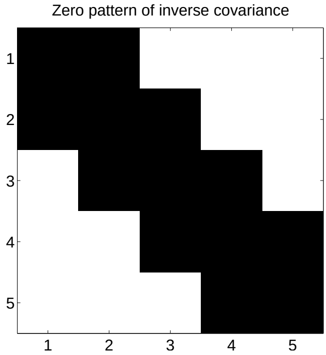


- factorization based on clique structure of graph:

$$\mathbb{P}(x_1, \ldots, x_p; \theta) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{C \in \mathcal{C}} \theta_C(x_C) \Big\}$$

Gaussian graphical models



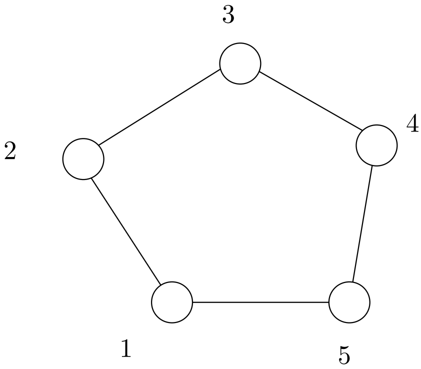


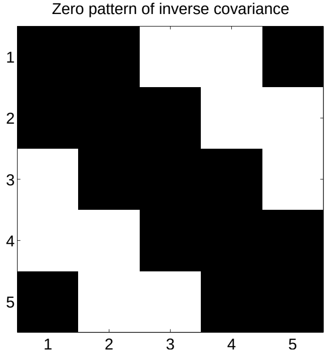
- density of multivariate Gaussian $X \sim N(0, \Theta^{-1})$:

$$\mathbb{P}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\big(-\frac{1}{2} x^T \Theta x\big).$$

- graphical model structured specified by zero-pattern of *inverse covariance* matrix Θ

Gaussian graphical models



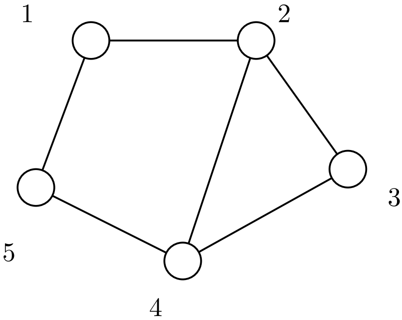


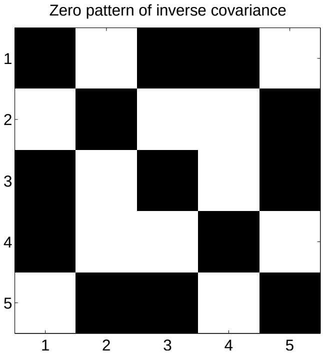
- density of multivariate Gaussian $X \sim N(0, \Theta^{-1})$:

$$\mathbb{P}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\big(-\frac{1}{2} x^T \Theta x\big).$$

- graphical model structured specified by zero-pattern of *inverse covariance* matrix Θ

Gaussian graphical models





- density of multivariate Gaussian $X \sim N(0, \Theta^{-1})$:

$$\mathbb{P}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\big(-\frac{1}{2} x^T \Theta x\big).$$

- graphical model structured specified by zero-pattern of *inverse covariance* matrix Θ

Discrete-variable graphical models

- random variable X_s at node s takes values in discrete space (e.g., $\mathcal{X} = \{-1, +1\}$)
- hierarchies of probability distributions:

Discrete-variable graphical models

- random variable X_s at node s takes values in discrete space (e.g., $\mathcal{X} = \{-1, +1\}$)
- hierarchies of probability distributions:
  - Independence model:

$$\mathbb{P}(x) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s \Big\}.$$

Discrete-variable graphical models

- random variable X_s at node s takes values in discrete space (e.g., $\mathcal{X} = \{-1, +1\}$)
- hierarchies of probability distributions:
  - Independence model:

$$\mathbb{P}(x) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s \Big\}.$$

  - Pairwise MRF (Ising model, 1923)

$$\mathbb{P}(x) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}.$$

Discrete-variable graphical models

- random variable X_s at node s takes values in discrete space (e.g., $\mathcal{X} = \{-1, +1\}$)
- hierarchies of probability distributions:
  - Independence model:

$$\mathbb{P}(x) = \frac{1}{Z(\theta)} \exp\Big\{\sum_{s \in V} \theta_s x_s\Big\}.$$

  - Pairwise MRF (Ising model, 1923)

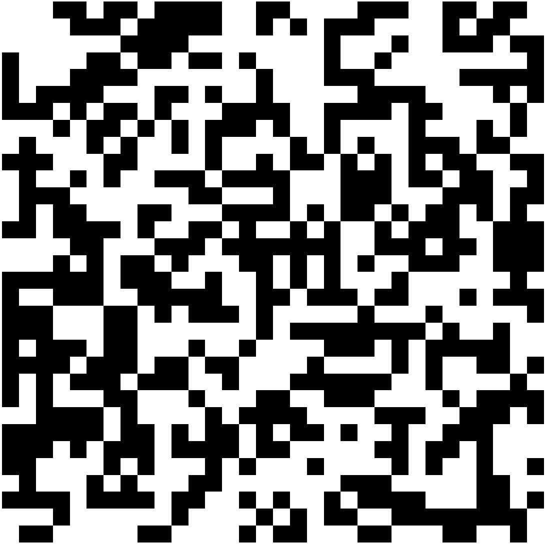
$$\mathbb{P}(x) = \frac{1}{Z(\theta)} \exp\Big\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\Big\}.$$

  - Triplet MRF

$$\mathbb{P}(x) = \frac{1}{Z(\theta)} \exp\Big\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E_2} \theta_{st} x_s x_t + \sum_{(s,t,u) \in E_3} \theta_{stu} x_s x_t x_u\Big\}.$$

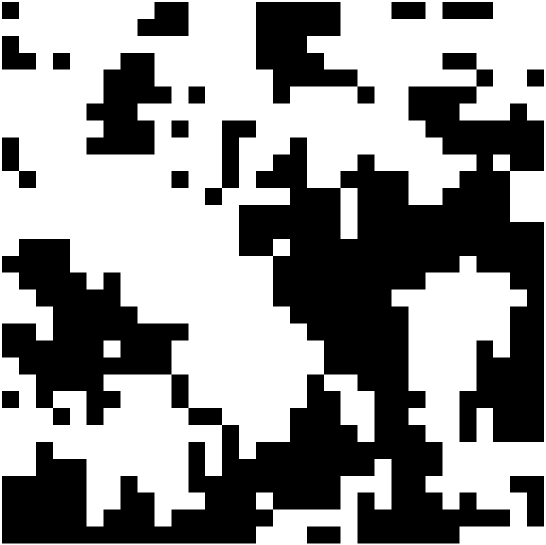
- (hyper)graph structure enforces that $\theta_{uv} = 0$ for all $(uv) \notin E$

Samples from binary-valued pairwise MRFs



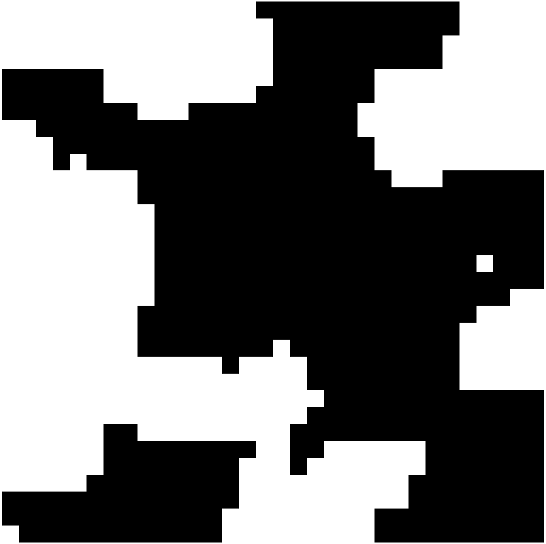
Independence model $\theta_{st} = 0$

Samples from binary-valued pairwise MRFs



Medium coupling $\theta_{st} \approx 0.2$

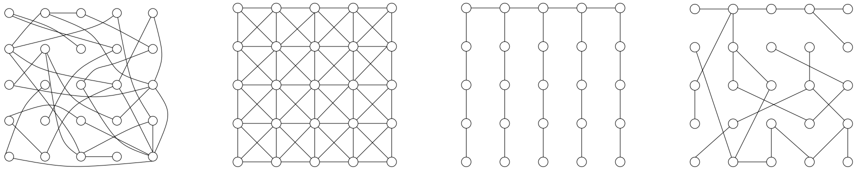
Samples from binary-valued pairwise MRFs



Strong coupling $\theta_{st} \approx 0.8$

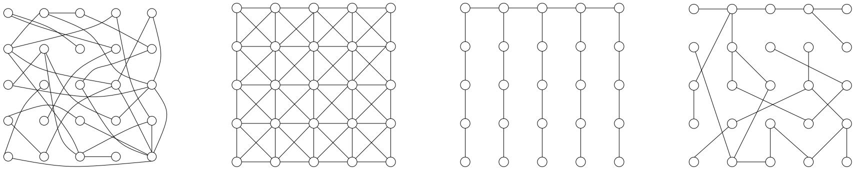
Graphical model selection

- let $G = (V, E)$ be an undirected graph on $p = |V|$ vertices



Graphical model selection

- let $G = (V, E)$ be an undirected graph on $p = |V|$ vertices

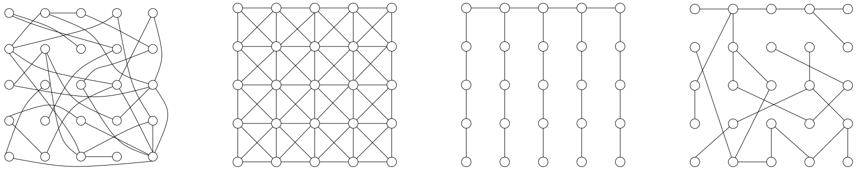


- pairwise Markov random field: family of prob. distributions

$$\mathbb{P}(x_1, \ldots, x_p; \theta) \quad = \quad \frac{1}{Z(\theta)} \exp\Big\{ \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

Graphical model selection

- let $G = (V, E)$ be an undirected graph on $p = |V|$ vertices



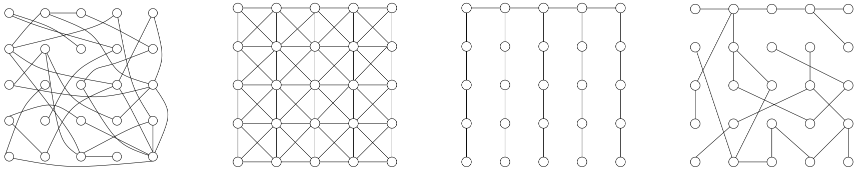
- pairwise Markov random field: family of prob. distributions

$$\mathbb{P}(x_1, \ldots, x_p; \theta) \quad = \quad \frac{1}{Z(\theta)} \exp\Big\{ \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- Problem of graph selection: given n independent and identically distributed (i.i.d.) samples of $X = (X_1, \ldots, X_p)$, identify the underlying graph structure

Graphical model selection

- let $G = (V, E)$ be an undirected graph on $p = |V|$ vertices

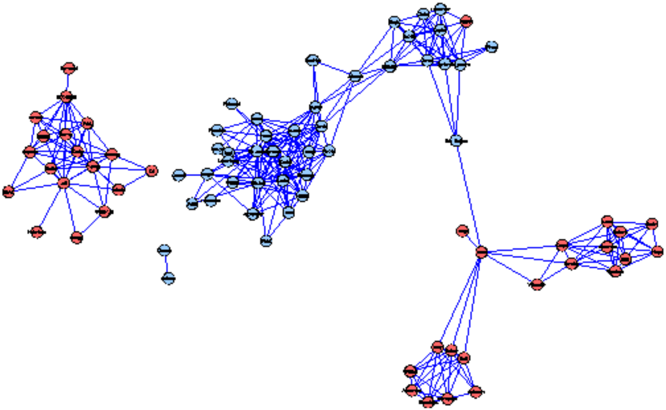


- pairwise Markov random field: family of prob. distributions

$$\mathbb{P}(x_1, \ldots, x_p; \theta) \quad = \quad \frac{1}{Z(\theta)} \exp\Big\{ \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- Problem of graph selection: given n independent and identically distributed (i.i.d.) samples of $X = (X_1, \ldots, X_p)$, identify the underlying graph structure
- complexity constraint: restrict to subset $\mathcal{G}_{d,p}$ of graphs with maximum degree d

Illustration: Voting behavior of US senators



Graphical model fit to voting records of US senators (Bannerjee, El Ghaoui, & d'Aspremont, 2008)

Outline of remainder of talk

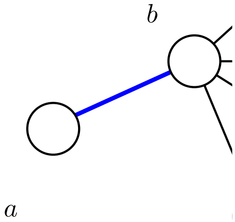
1. Background and framework
   (a) Problem set-up
   (b) Some challenges in distinguishing graphs
   (c) Analysis in a high-dimensional framework

2. Practical schemes
   (a) Gaussian graphical models via log-determinant
   (b) Discrete graphical models via logistic regression
   (c) Sufficient conditions for high-dimensional consistency

3. Fundamental limits
   (a) An unorthodox channel coding problem
   (b) Necessary conditions
   (c) Sufficient conditions (optimal algorithms)

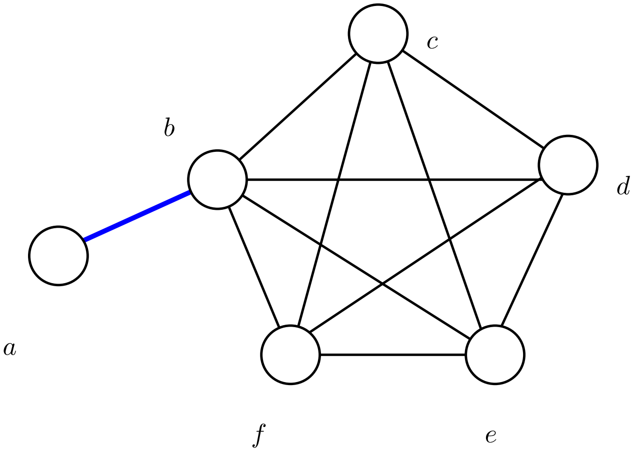
4. Various open questions......

Guilt by association



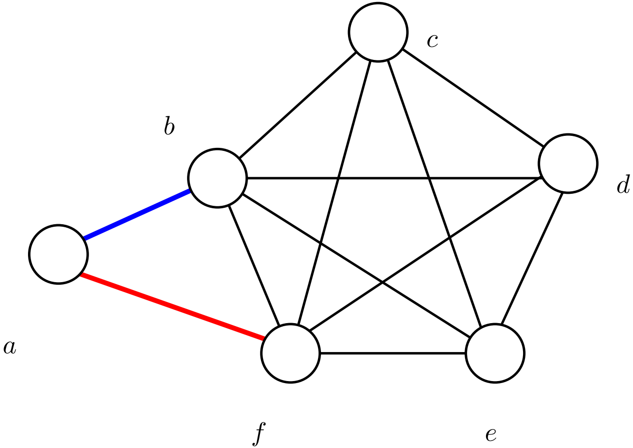
- Andrew (a) and Bob (b) are brothers

Guilt by association



- Andrew (a) and Bob (b) are brothers
- Bob (b) is part of a criminal network

Guilt by association



- Andrew (a) and Bob (b) are brothers
- Bob (b) is part of a criminal network
- Is Andrew also a criminal?

Some challenges in distinguishing graphs

- clearly, a lower bound on the minimum edge weight is required:

$$\min_{(s,t)\in E} |\theta^*_{st}| \geq \theta_{\min},$$

  although $\theta_{\min}(p, d) = o(1)$ is allowed.

- in contrast to other testing/detection problems, large $|\theta_{st}|$ also problematic

Some challenges in distinguishing graphs

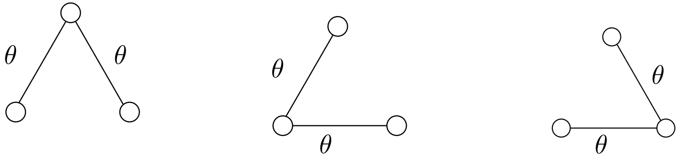
- clearly, a lower bound on the minimum edge weight is required:

$$\min_{(s,t)\in E} |\theta^*_{st}| \geq \theta_{\min},$$

although $\theta_{\min}(p,d) = o(1)$ is allowed.

- in contrast to other testing/detection problems, large $|\theta_{st}|$ also problematic

Toy example: Graphs from $\mathcal{G}_{3,2}$ (i.e., $p = 3$; $d = 2$), and $x \in \{-1,+1\}^3$

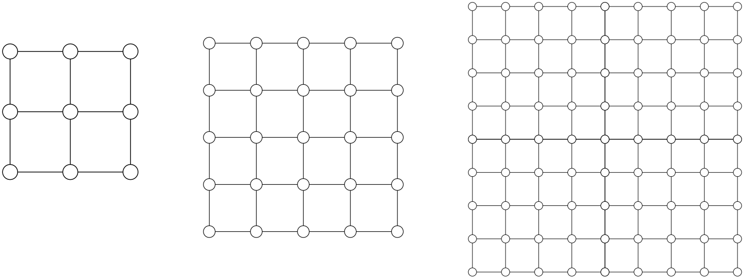


As θ increases, all three Markov random fields become arbitrarily close to:

$$\mathbb{P}(x_1, x_2, x_3) = \begin{cases} 1/2 & \text{if } x \in \{(-1)^3, (+1)^3\} \\ 0 & \text{otherwise.} \end{cases}$$

High-dimensional analysis

- classical analysis: dimension p fixed, sample size $n \to +\infty$
- high-dimensional analysis: allow both dimension p, sample size n, and maximum degree d to increase at arbitrary rates



- take n i.i.d. samples from MRF defined by $G_{p,d}$
- study probability of success as a function of three parameters:

$$\text{Success}(n, p, d) \quad = \quad \mathbb{P}[\text{Method recovers graph } G_{p,d} \text{ from } n \text{ samples}]$$

- theory is non-asymptotic: explicit probabilities for finite (n, p, d)

Some issues in graph selection

Consider some fixed loss function, and a fixed level δ of error.

Limitations of tractable algorithms:
Given particular (polynomial-time) algorithms

- for what sample sizes n do they succeed/fail to achieve error δ?
- given a collection of methods, when does more computation reduce minimum # samples needed?

Information-theoretic limitations:
Data collection as communication from nature $\longrightarrow$ statistician:

- what are fundamental limitations of problem (Shannon capacity)?
- when are known (polynomial-time) methods optimal?
- when are there gaps between poly.-time methods and optimal methods?

§2. Practical methods: Gaussian graphical selection

- recall form of Gaussian density in terms of *inverse covariance* Θ:

$$\mathbb{P}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\big(-\frac{1}{2} x^T \Theta x\big).$$

§2. Practical methods: Gaussian graphical selection

- recall form of Gaussian density in terms of *inverse covariance* Θ:

$$\mathbb{P}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\big(-\frac{1}{2} x^T \Theta x\big).$$

ℓ_1-regularized maximum likelihood:

$$\widehat{\Theta} \quad = \quad \arg\min_{\Theta \succ 0} \Big\{ \underbrace{-\log\det\Theta + \langle\langle \widehat{\Sigma}^n, \; \Theta \rangle\rangle}_{\text{neg. log likelihood}} \quad + \quad \underbrace{\rho_n \sum_{i \neq j} |\Theta_{ij}|}_{\text{regularizer}} \Big\}.$$

§2. Practical methods: Gaussian graphical selection

- recall form of Gaussian density in terms of *inverse covariance* Θ:

$$\mathbb{P}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\big(-\frac{1}{2} x^T \Theta x\big).$$

ℓ_1-regularized maximum likelihood:

$$\widehat{\Theta} \quad = \quad \arg\min_{\Theta \succ 0} \Big\{ \underbrace{-\log\det\Theta + \langle\!\langle \widehat{\Sigma}^n, \; \Theta \rangle\!\rangle}_{\text{neg. log likelihood}} \quad + \quad \underbrace{\rho_n \sum_{i \neq j} |\Theta_{ij}|}_{\text{regularizer}} \Big\}.$$

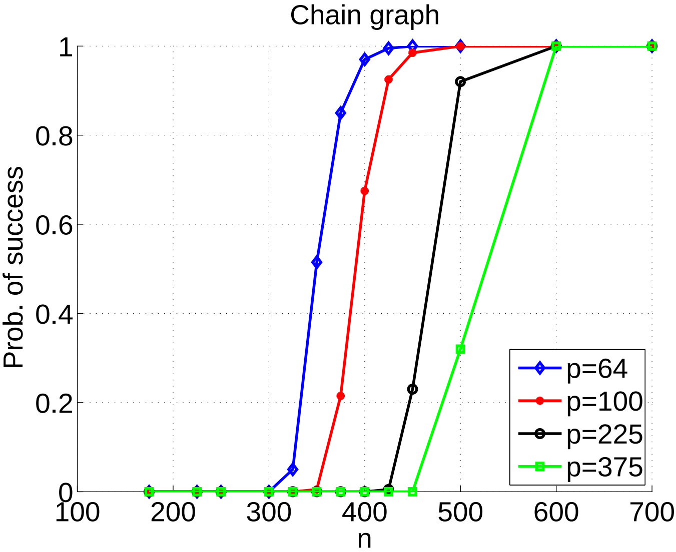
- plug-in approach: use samples $X^{(k)}, k = 1, \ldots, n$ to estimate covariance matrix $\Sigma = \Theta^{-1}$ via the sample covariance

$$\widehat{\Sigma}^n \quad := \quad \frac{1}{n} \sum_{k=1}^{n} X^{(k)} (X^{(k)})^T.$$

- regularization parameter $\rho_n > 0$ is a user-specified quantity

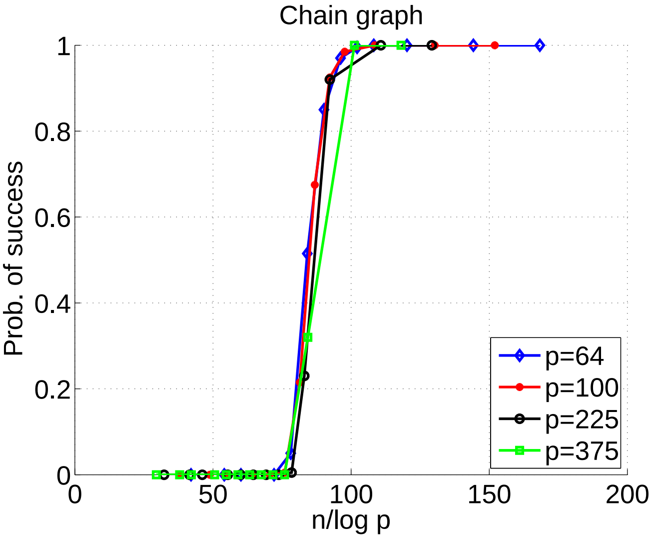
(e.g., Yuan & Lin, 2006; d'Asprémont et al., 2007; Friedman, 2008; Rothman et al., 2008)

Empirical behavior: Unrescaled plots



Plots of success probability versus raw sample size n.

Empirical behavior: Appropriately rescaled



Plots of success probability versus control parameter $\theta_{LR}(n, p, d)$.

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant
  $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem

For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

and regularization parameter $\rho_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$*, then with probability greater than* $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$*:*

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem

For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

and regularization parameter $\rho_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$*, then with probability greater than* $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$*:*

(a) *No false inclusions: The regularized log-determinant estimate* $\widehat{\Theta}$ *returns an edge set* $\widehat{E} \subseteq E^*$*.*

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem

For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

and regularization parameter $\rho_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$*, then with probability greater than* $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$*:*

(a) *No false inclusions: The regularized log-determinant estimate* $\widehat{\Theta}$ *returns an edge set* $\widehat{E} \subseteq E^*$*.*

(b) ℓ_∞*-control: Estimate satisfies* $\max_{i,j} |\widehat{\Theta}_{ij} - \Theta^*_{ij}| \leq 2\, c_4 \sqrt{\frac{\tau \log p}{n}}$.

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem

For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

and regularization parameter $\rho_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$*, then with probability greater than* $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$*:*

(a) *No false inclusions: The regularized log-determinant estimate* $\widehat{\Theta}$ *returns an edge set* $\widehat{E} \subseteq E^*$*.*

(b) *ℓ_∞-control: Estimate satisfies* $\max_{i,j} |\widehat{\Theta}_{ij} - \Theta^*_{ij}| \leq 2\, c_4 \sqrt{\frac{\tau \log p}{n}}$.

(c) *Model selection consistency: If* $\theta_\star \geq c_4 \sqrt{\frac{\tau \log p}{n}}$*, then* $E = \widehat{E}$*.*

Some consequences

Corollary

Under same conditions as theorem, operator norm consistency at rates:

$$\max\left\{ |||\widehat{\Theta} - \Theta^*|||_2, |||\widehat{\Sigma} - \Sigma^*|||_2 \right\} \quad = \quad \mathcal{O}\big(\sqrt{\frac{d^2 \log p}{n}}\big),$$

Some consequences

Corollary

Under same conditions as theorem, operator norm consistency at rates:

$$\max\left\{|||\widehat{\Theta}-\Theta^*|||_2, |||\widehat{\Sigma}-\Sigma^*|||_2\right\} \quad = \quad \mathcal{O}\big(\sqrt{\frac{d^2\log p}{n}}\big),$$

- sample covariance estimate is highly inconsistent in this regime:

$$|||\frac{1}{n}\sum_{i=1}^n X^{(i)}(X^{(i)})^T - \Sigma|||_2 \geq c\sqrt{\frac{p}{n}} \;\rightarrow\; +\infty$$

(Marcenko & Pastur, 1967; Davidson & Szarek, 2001)

Some consequences

Corollary

Under same conditions as theorem, operator norm consistency at rates:

$$\max\{\|\widehat{\Theta} - \Theta^*\|_2, \|\widehat{\Sigma} - \Sigma^*\|_2\} \quad = \quad \mathcal{O}(\sqrt{\frac{d^2 \log p}{n}}),$$

- sample covariance estimate is highly inconsistent in this regime:

$$\|\frac{1}{n}\sum_{i=1}^{n} X^{(i)}(X^{(i)})^T - \Sigma\|_2 \geq c\sqrt{\frac{p}{n}} \rightarrow +\infty$$

(Marcenko & Pastur, 1967; Davidson & Szarek, 2001)

- Rothman et al. (2008) showed

$$\max\{\|\widehat{\Theta} - \Theta^*\|_2, \|\widehat{\Sigma} - \Sigma^*\|_2\} = \mathcal{O}(\sqrt{\frac{s \log p}{n}}), \quad \text{where } s = \#\text{ edges},$$

which is substantially weaker for d-regular graphs with $s = \Theta(dp)$

Global max. likelihood for discrete models?

- maximum likelihood for general graphical model in exponential family:

$$\widehat{\theta} \quad = \quad \arg\max_{\theta \in \mathbb{R}^p} \left\{ \sum_{(s,t) \in E} \theta_{st} \underbrace{\widehat{\mathbb{E}}[X_s X_t]}_{\text{empirical moments}} - \log Z(\theta) \right\}$$

Global max. likelihood for discrete models?

- maximum likelihood for general graphical model in exponential family:

$$\widehat{\theta} \quad = \quad \arg\max_{\theta \in \mathbb{R}^p} \left\{ \sum_{(s,t) \in E} \theta_{st} \underbrace{\widehat{\mathbb{E}}[X_s X_t]}_{\text{empirical moments}} - \log Z(\theta) \right\}$$

- exact likelihood involves log partition function

$$\log Z(\theta) = \begin{cases} \int_{\mathbb{R}^n} \exp(-\frac{1}{2} x^T \Theta x) dx & \text{for Gaussian RV} \\ \sum_{x \in -1,+1^p} \exp\left(\sum_{(s,t)} \theta_{st} x_s x_t \right) & \text{for binary RV} \end{cases}$$

Global max. likelihood for discrete models?

- maximum likelihood for general graphical model in exponential family:

$$\widehat{\theta} \quad = \quad \arg\max_{\theta \in \mathbb{R}^p} \Big\{ \sum_{(s,t) \in E} \theta_{st} \underbrace{\widehat{\mathbb{E}}[X_s X_t]}_{\text{empirical moments}} - \log Z(\theta) \Big\}$$

- exact likelihood involves log partition function

$$\log Z(\theta) = \begin{cases} \int_{\mathbb{R}^n} \exp(-\frac{1}{2} x^T \Theta x) dx & \text{for Gaussian RV} \\ \sum_{x \in -1,+1^p} \exp\big(\sum_{(s,t)} \theta_{st} x_s x_t \big) & \text{for binary RV} \end{cases}$$

- key consequence: likelihood computation is
  - straightfoward for Gaussian MRFs (log-determinant)
  - intractable for Ising models (binary pairwise MRFs) (Welsh, 1993)

Global max. likelihood for discrete models?

- maximum likelihood for general graphical model in exponential family:

$$\widehat{\theta} \quad = \quad \arg\max_{\theta \in \mathbb{R}^p} \left\{ \sum_{(s,t) \in E} \theta_{st} \underbrace{\widehat{\mathbb{E}}[X_s X_t]}_{\text{empirical moments}} - \log Z(\theta) \right\}$$

- exact likelihood involves log partition function

$$\log Z(\theta) = \begin{cases} \int_{\mathbb{R}^n} \exp(-\frac{1}{2} x^T \Theta x) dx & \text{for Gaussian RV} \\ \sum_{x \in -1, +1^p} \exp\left(\sum_{(s,t)} \theta_{st} x_s x_t \right) & \text{for binary RV} \end{cases}$$

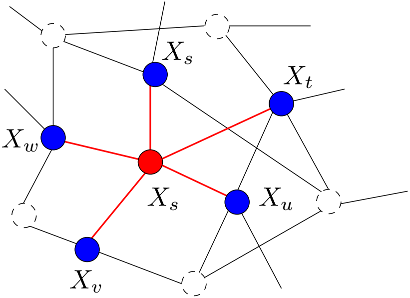
- key consequence: likelihood computation is
  - straightfoward for Gaussian MRFs (log-determinant)
  - intractable for Ising models (binary pairwise MRFs) (Welsh, 1993)
- possible work-arounds:
  - MCMC methods
  - stochastic approximation methods
  - variational approximations (mean field, Bethe and belief propagation)

Markov property and neighborhood structure

- Markov properties encode neighborhood structure:

$$\underbrace{(X_s \mid X_{V\backslash s})}_{\text{Condition on full graph}} \stackrel{d}{=} \underbrace{(X_s \mid X_{N(s)})}_{\text{Condition on Markov blanket}}$$

$N(s) = \{s, t, u, v, w\}$



- basis of pseudolikelihood method (Besag, 1974)
- used for Gaussian model selection (Meinshausen & Buhlmann, 2006)

Graph selection via neighborhood regression

Observation: Recovering graph G equivalent to recovering neighborhood set $N(s)$ for all $s \in V$.

Method: Given n i.i.d. samples $\{X^{(1)}, \ldots, X^{(n)}\}$, perform logistic regression of each node X_s on $X_{\backslash s} := \{X_s, \; t \neq s\}$ to estimate neighborhood structure $\widehat{N}(s)$.

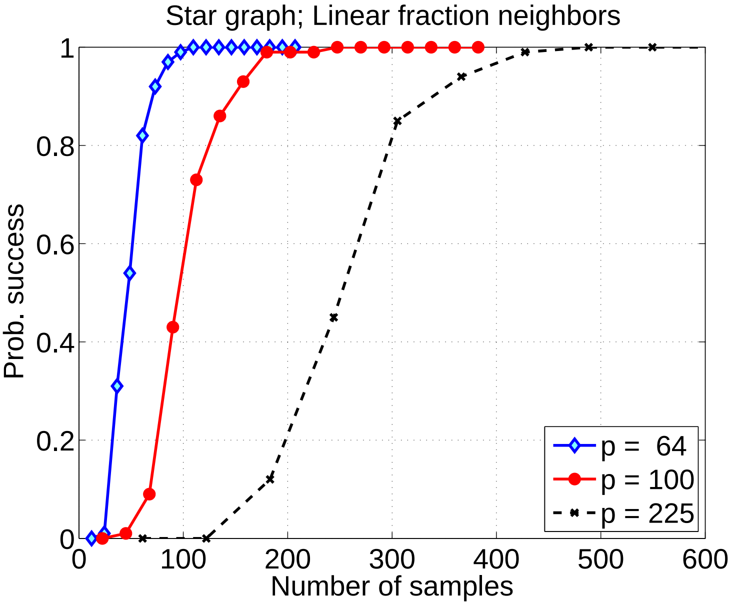
1. For each node $s \in V$, perform ℓ_1 regularized logistic regression of X_s on the remaining variables $X_{\backslash s}$:

$$\widehat{\theta}[s] \quad := \quad \arg \min_{\theta \in \mathbb{R}^{p-1}} \Big\{ \quad \underbrace{\frac{1}{n} \sum_{i=1}^{n} f(\theta; X_{\backslash s}^{(i)})}_{\text{logistic likelihood}} \quad + \quad \underbrace{\rho_n \|\theta\|_1}_{\text{regularization}} \Big\}$$

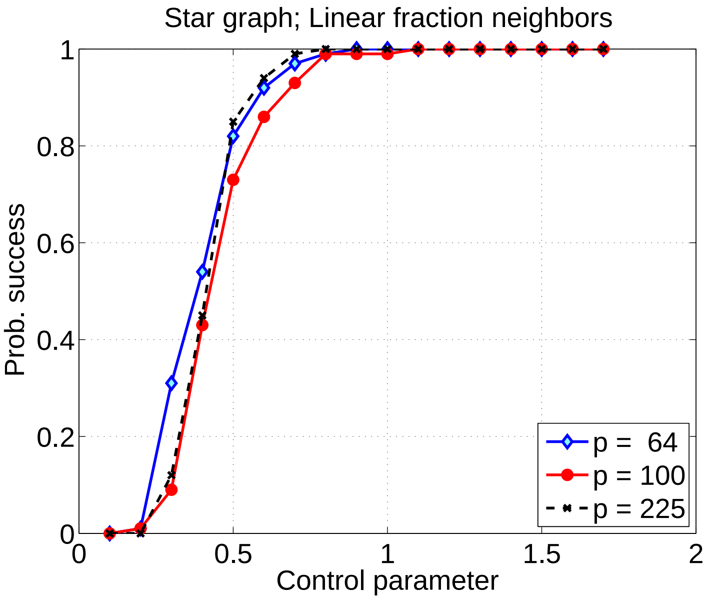
2. Estimate the local neighborhood $\widehat{N}(s)$ as the support (non-negative entries) of the regression vector $\widehat{\theta}[s]$.

3. Combine the neighborhood estimates in a consistent manner (AND, or OR rule).

Empirical behavior: Unrescaled plots



Empirical behavior: Appropriately rescaled



Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- edge weights $|\theta_{st}| \geq \theta_{\min}$ for all $(s, t) \in E$
- draw n i.i.d, samples, and analyze prob. success indexed by (n, p, d)

Theorem

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- edge weights $|\theta_{st}| \geq \theta_{\min}$ for all $(s,t) \in E$
- draw n i.i.d, samples, and analyze prob. success indexed by (n, p, d)

Theorem

Under incoherence conditions, for a rescaled sample size *(RavWaiLaf06)*

$$\theta_{LR}(n,p,d) \quad := \quad \frac{n}{d^3 \log p} \; > \; \theta_{\text{crit}}$$

and regularization parameter $\rho_n \geq c_1 \, \tau \sqrt{\frac{\log p}{n}}$, *then with probability greater than* $1 - 2\exp\big(-c_2(\tau - 2)\log p\big) \rightarrow 1$*:*

(a) Uniqueness: *For each node* $s \in V$, *the* ℓ_1*-regularized logistic convex program has a unique solution.* *(Non-trivial since* $p \gg n \implies$ *not strictly convex).*

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- edge weights $|\theta_{st}| \geq \theta_{\min}$ for all $(s,t) \in E$
- draw n i.i.d, samples, and analyze prob. success indexed by (n, p, d)

Theorem

Under incoherence conditions, for a rescaled sample size *(RavWaiLaf06)*

$$\theta_{LR}(n,p,d) \quad := \quad \frac{n}{d^3 \log p} \; > \; \theta_{\text{crit}}$$

and regularization parameter $\rho_n \geq c_1 \, \tau \sqrt{\frac{\log p}{n}}$, *then with probability greater than* $1 - 2\exp\big(-c_2(\tau - 2)\log p\big) \rightarrow 1$:

(a) **Uniqueness:** *For each node* $s \in V$, *the* ℓ_1*-regularized logistic convex program has a unique solution.* *(Non-trivial since* $p \gg n \implies$ *not strictly convex).*

(b) **Correct exclusion:** *The estimated sign neighborhood* $\widehat{N}(s)$ *correctly excludes all edges not in the true neighborhood.*

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- edge weights $|\theta_{st}| \geq \theta_{\min}$ for all $(s,t) \in E$
- draw n i.i.d, samples, and analyze prob. success indexed by (n, p, d)

Theorem

Under incoherence conditions, for a rescaled sample size *(RavWaiLaf06)*

$$\theta_{LR}(n,p,d) \quad := \quad \frac{n}{d^3 \log p} \; > \; \theta_{\text{crit}}$$

and regularization parameter $\rho_n \geq c_1 \, \tau \sqrt{\frac{\log p}{n}}$, *then with probability greater than* $1 - 2\exp\big(-c_2(\tau - 2)\log p\big) \rightarrow 1$*:*

(a) **Uniqueness:** *For each node* $s \in V$*, the* ℓ_1*-regularized logistic convex program has a unique solution.* *(Non-trivial since* $p \gg n \implies$ *not strictly convex).*

(b) **Correct exclusion:** *The estimated sign neighborhood* $\widehat{N}(s)$ *correctly excludes all edges not in the true neighborhood.*

(c) **Correct inclusion:** *For* $\theta_{\min} \geq c_3 \tau \sqrt{d} \rho_n$*, the method selects the correct signed neighborhood.*

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- edge weights $|\theta_{st}| \geq \theta_{\min}$ for all $(s,t) \in E$
- draw n i.i.d, samples, and analyze prob. success indexed by (n, p, d)

Theorem

Under incoherence conditions, for a rescaled sample size *(RavWaiLaf06)*

$$\theta_{LR}(n,p,d) \quad := \quad \frac{n}{d^3 \log p} \; > \; \theta_{\text{crit}}$$

and regularization parameter $\rho_n \geq c_1 \, \tau \sqrt{\frac{\log p}{n}}$, *then with probability greater than* $1 - 2\exp\big(-c_2(\tau - 2)\log p\big) \rightarrow 1$*:*

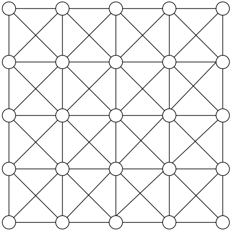
(a) Uniqueness: *For each node* $s \in V$, *the* ℓ_1*-regularized logistic convex program has a unique solution. (Non-trivial since* $p \gg n \implies$ *not strictly convex).*

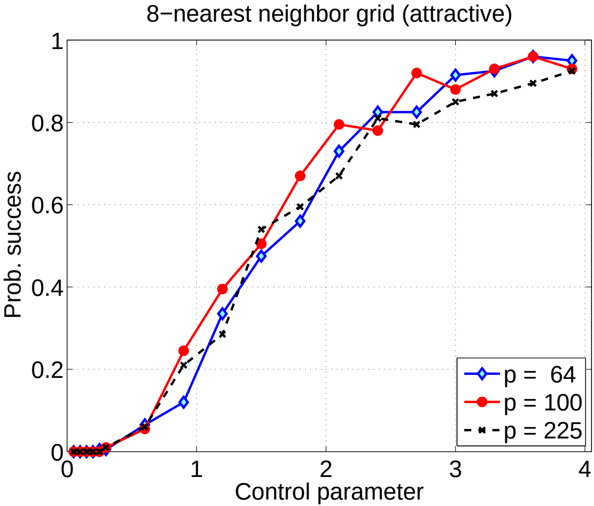
(b) Correct exclusion: *The estimated sign neighborhood* $\widehat{N}(s)$ *correctly excludes all edges not in the true neighborhood.*

(c) Correct inclusion: *For* $\theta_{\min} \geq \; c_3 \tau \sqrt{d} \rho_n$, *the method selects the correct signed neighborhood.*

Consequence: For $\theta_{\min} = \Omega(1/d)$, it suffices to have $n = \Omega(d^3 \log p)$.

Results for 8-grid graphs





Prob. of success $\mathbb{P}[\widehat{G} = G]$ versus rescaled sample size $\theta_{LR}(n, p, d^3) = \frac{n}{d^3 \log p}$

Assumptions

Define Fisher information matrix of logistic regression:
$Q^* := \mathbb{E}_{\theta^*}\left[\nabla^2 f(\theta^*; X)\right]$.

A1. Dependency condition: Bounded eigenspectra:

$$C_{min} \;\leq\; \lambda_{min}(Q^*_{SS}), \qquad \text{and} \qquad \lambda_{max}(Q^*_{SS}) \;\leq\; C_{max}.$$
$$\lambda_{max}(\mathbb{E}_{\theta^*}[XX^T]) \quad \leq \quad D_{\max}.$$

A2. Incoherence There exists an $\nu \in (0, 1]$ such that

$$|||Q^*_{S^cS}(Q^*_{SS})^{-1}|||_{\infty,\infty} \quad \leq \quad 1 - \nu.$$

where $|||A|||_{\infty,\infty} := \max_i \sum_j |A_{ij}|$.

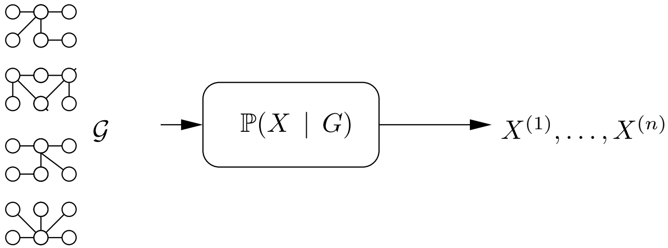
- bounds on eigenvalues are fairly standard
- incoherence condition:
  - partly necessary (prevention of degenerate models)
  - partly an artifact of ℓ_1-regularization
- incoherence condition is weaker than correlation decay

§3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:

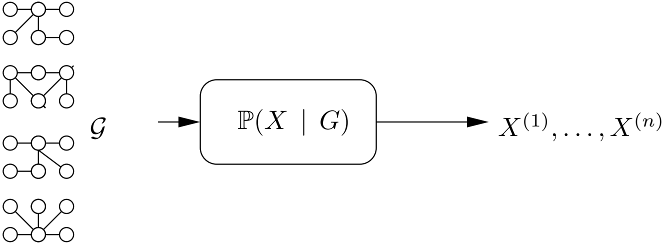
§3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
  - codewords/codebook: graph G in some graph class $\mathcal{G}$
  - channel use: draw sample $X^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})$ from Markov random field $\mathbb{P}_{\theta(G)}$
  - decoding problem: use n samples $\{X^{(1)}, \ldots, X^{(n)}\}$ to correctly distinguish the "codeword"



§3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
  - codewords/codebook: graph G in some graph class $\mathcal{G}$
  - channel use: draw sample $X^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})$ from Markov random field $\mathbb{P}_{\theta(G)}$
  - decoding problem: use n samples $\{X^{(1)}, \ldots, X^{(n)}\}$ to correctly distinguish the "codeword"



Channel capacity for graph decoding determined by balance between

- log number of models
- relative distinguishability of different models

Necessary conditions for $\mathcal{G}_{d,p}$

- $G \in \mathcal{G}_{d,p}$: graphs with p nodes and max. degree d
- Ising models with:
  - *Minimum edge weight:* $|\theta^*_{st}| \geq \theta_{\min}$ for all edges
  - *Maximum neighborhood weight:* $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta^*_{st}|$

Necessary conditions for $\mathcal{G}_{d,p}$

- $G \in \mathcal{G}_{d,p}$: graphs with p nodes and max. degree d
- Ising models with:
  - *Minimum edge weight:* $|\theta^*_{st}| \geq \theta_{\min}$ for all edges
  - *Maximum neighborhood weight:* $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta^*_{st}|$

Theorem

If the sample size n is upper bounded by *(Santhanam & W, 2008)*

$$n \quad < \quad \max\Big\{\frac{d}{8}\log\frac{p}{8d}, \; \frac{\exp(\frac{\omega(\theta)}{4})\, d\theta_{\min}\log(pd/8)}{128\exp(\frac{3\theta_{\min}}{2})}, \; \frac{\log p}{2\theta_{\min}\tanh(\theta_{\min})}\Big\}$$

then the probability of error of any algorithm over $\mathcal{G}_{d,p}$ is at least $1/2$.

Necessary conditions for $\mathcal{G}_{d,p}$

- $G \in \mathcal{G}_{d,p}$: graphs with p nodes and max. degree d
- Ising models with:
  - *Minimum edge weight:* $|\theta^*_{st}| \geq \theta_{\min}$ for all edges
  - *Maximum neighborhood weight:* $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta^*_{st}|$

Theorem

If the sample size n is upper bounded by *(Santhanam & W, 2008)*

$$n \quad < \quad \max\left\{ \frac{d}{8} \log \frac{p}{8d}, \; \frac{\exp(\frac{\omega(\theta)}{4})\, d\theta_{\min} \log(pd/8)}{128 \exp(\frac{3\theta_{\min}}{2})}, \; \frac{\log p}{2\theta_{\min} \tanh(\theta_{\min})} \right\}$$

then the probability of error of any algorithm over $\mathcal{G}_{d,p}$ is at least $1/2$.

Interpretation:

- Naive bulk effect: Arises from log cardinality $\log |\mathcal{G}_{d,p}|$
- d-clique effect: Difficulty of separating models that contain a near d-clique
- Small weight effect: Difficult to detect edges with small weights.

Some consequences

Corollary

For asymptotically reliable recovery over $\mathcal{G}_{d,p}$, any algorithm requires at least $n = \Omega(d^2 \log p)$ samples.

Some consequences

Corollary

For asymptotically reliable recovery over $\mathcal{G}_{d,p}$, any algorithm requires at least $n = \Omega(d^2 \log p)$ samples.

- note that maximum neighborhood weight $\omega(\theta^*) \geq d\,\theta_{\min} \Longrightarrow$ require $\theta_{\min} = \mathcal{O}(1/d)$

Some consequences

Corollary

For asymptotically reliable recovery over $\mathcal{G}_{d,p}$, any algorithm requires at least $n = \Omega(d^2 \log p)$ samples.

- note that maximum neighborhood weight $\omega(\theta^*) \geq d\,\theta_{\min} \Longrightarrow$ require $\theta_{\min} = \mathcal{O}(1/d)$
- from small weight effect

$$n = \Omega\Big(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min})}\Big) \;=\; \Omega\Big(\frac{\log p}{\theta_{\min}^2}\Big)$$

Some consequences

Corollary

For asymptotically reliable recovery over $\mathcal{G}_{d,p}$, any algorithm requires at least $n = \Omega(d^2 \log p)$ samples.

- note that maximum neighborhood weight $\omega(\theta^*) \geq d\,\theta_{\min} \implies$ require $\theta_{\min} = \mathcal{O}(1/d)$
- from small weight effect

$$n = \Omega\Big(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min})}\Big) \;=\; \Omega\Big(\frac{\log p}{\theta_{\min}^2}\Big)$$

- conclude that ℓ_1-regularized logistic regression (LR) is within $\Theta(d)$ of optimal for general graphs (Ravikumar., W. & Lafferty, 2006)

Some consequences

Corollary

For asymptotically reliable recovery over $\mathcal{G}_{d,p}$, any algorithm requires at least $n = \Omega(d^2 \log p)$ samples.

- note that maximum neighborhood weight $\omega(\theta^*) \geq d\,\theta_{\min} \Longrightarrow$ require $\theta_{\min} = \mathcal{O}(1/d)$
- from small weight effect

$$n = \Omega\Big(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min})}\Big) \;=\; \Omega\Big(\frac{\log p}{\theta_{\min}^2}\Big)$$

- conclude that ℓ_1-regularized logistic regression (LR) is within $\Theta(d)$ of optimal for general graphs (Ravikumar., W. & Lafferty, 2006)
- for bounded degree graphs:
  - ℓ_1-LR order-optimal under incoherence conditions with cost $\mathcal{O}(p^4)$
  - thresholding procedure order-optimal under correlation decay, also with polynomial complexity (Bresler, Sly & Mossel, 2008)

Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles $\mathcal{G} \subseteq \mathcal{G}_{p,d}$

Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
- choose $G \in \mathcal{G}$ u.a.r., and consider multi-way hypothesis testing problem based on the data $\mathbf{X}_1^n = \{X^{(1)}, \ldots, X^{(n)}\}$

Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
- choose $G \in \mathcal{G}$ u.a.r., and consider multi-way hypothesis testing problem based on the data $\mathbf{X}_1^n = \{X^{(1)}, \ldots, X^{(n)}\}$
- for any graph estimator $\psi : \mathcal{X}^n \rightarrow \mathcal{G}$, Fano's inequality implies that

$$\mathbb{P}[\psi(\mathbf{X}_1^n) \neq G] \geq 1 - \frac{I(\mathbf{X}_1^n; G)}{\log |\mathcal{G}|} - o(1)$$

  where $I(\mathbf{X}_1^n; G)$ is mutual information between observations $\mathbf{X}_1^n$ and randomly chosen graph G

Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
- choose $G \in \mathcal{G}$ u.a.r., and consider multi-way hypothesis testing problem based on the data $\mathbf{X}_1^n = \{X^{(1)}, \ldots, X^{(n)}\}$
- for any graph estimator $\psi : \mathcal{X}^n \to \mathcal{G}$, Fano's inequality implies that

$$\mathbb{P}[\psi(\mathbf{X}_1^n) \neq G] \geq 1 - \frac{I(\mathbf{X}_1^n; G)}{\log |\mathcal{G}|} - o(1)$$

  where $I(\mathbf{X}_1^n; G)$ is mutual information between observations $\mathbf{X}_1^n$ and randomly chosen graph G

- remaining steps:

  1. Construct "difficult" sub-ensembles $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
  2. Compute or lower bound the log cardinality $\log |\mathcal{G}|$.
  3. Upper bound the mutual information $I(\mathbf{X}_1^n; G)$.

Two straightforward ensembles

Two straightforward ensembles

1. Naive bulk ensemble: All graphs on p vertices with max. degree d (i.e., $\mathcal{G} = \mathcal{G}_{p,d}$)

Two straightforward ensembles

1. Naive bulk ensemble: All graphs on p vertices with max. degree d (i.e., $\mathcal{G} = \mathcal{G}_{p,d}$)
   - simple counting argument: $\log|\mathcal{G}_{p,d}| = \Theta\big(pd\log(p/d)\big)$
   - trivial upper bound: $I(\mathbf{X}_1^n; G) \leq H(\mathbf{X}_1^n) \ \leq \ np$.
   - substituting into Fano yields necessary condition $n = \Omega(d\log(p/d))$
   - this bound independently derived by different approach by Bresler et al. (2008)

Two straightforward ensembles

1. Naive bulk ensemble: All graphs on p vertices with max. degree d (i.e., $\mathcal{G} = \mathcal{G}_{p,d}$)
   - simple counting argument: $\log|\mathcal{G}_{p,d}| = \Theta\big(pd\log(p/d)\big)$
   - trivial upper bound: $I(\mathbf{X}_1^n; G) \leq H(\mathbf{X}_1^n) \ \leq \ np.$
   - substituting into Fano yields necessary condition $n = \Omega(d\log(p/d))$
   - this bound independently derived by different approach by Bresler et al. (2008)

2. Small weight effect: Ensemble $\mathcal{G}$ consisting of graphs with a single edge with weight $\theta = \theta_{\min}$

Two straightforward ensembles

1. Naive bulk ensemble: All graphs on p vertices with max. degree d (i.e., $\mathcal{G} = \mathcal{G}_{p,d}$)
   - simple counting argument: $\log |\mathcal{G}_{p,d}| = \Theta\big(pd \log(p/d)\big)$
   - trivial upper bound: $I(\mathbf{X}_1^n; G) \leq H(\mathbf{X}_1^n) \;\leq\; np$.
   - substituting into Fano yields necessary condition $n = \Omega(d \log(p/d))$
   - this bound independently derived by different approach by Bresler et al. (2008)

2. Small weight effect: Ensemble $\mathcal{G}$ consisting of graphs with a single edge with weight $\theta = \theta_{\min}$
   - simple counting: $\log |\mathcal{G}| = \log \binom{p}{2}$
   - upper bound on mutual information:

$$I(\mathbf{X}_1^n; G) \leq \frac{1}{\binom{p}{2}} \sum_{(i,j),(k,\ell) \in E} D\big(\theta(G^{ij}) \| \theta(G^{k\ell})\big).$$

   - upper bound on symmetrized Kullback-Leibler divergences:

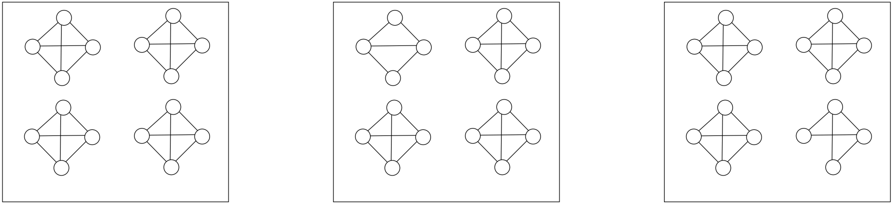
$$D\big(\theta(G^{ij}) \| \theta(G^{k\ell})\big) + D\big(\theta(G^{k\ell}) \| \theta(G^{ij})\big) \leq 2\theta_{\min} \tanh(\theta_{\min}/2)$$

   - substituting into Fano yields necessary condition $n = \Omega\big(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min}/2)}\big)$

A harder d-clique ensemble

Constructive procedure:

1. Divide the vertex set V into $\lfloor \frac{p}{d+1} \rfloor$ groups of size $d+1$.
2. Form the base graph $\bar{G}$ by making a $(d+1)$-clique within each group.
3. Form graph G^{uv} by deleting edge (u,v) from $\bar{G}$.
4. Form Markov random field $\mathbb{P}_{\theta(G^{uv})}$ by setting $\theta_{st} = \theta_{\min}$ for all edges.



(a) Base graph $\bar{G}$ (b) Graph G^{uv} (c) Graph G^{st}

- For $d \leq p/4$, we can form

$$|\mathcal{G}| \geq \lfloor \frac{p}{d+1} \rfloor \binom{d+1}{2} = \Omega(dp)$$

such graphs.

Summary and open questions

- *Practical methods:*
  - Log-determinant for Gaussian graphical models:

$$n > c_1 \ \max\{\frac{1}{\theta_{\min}^2}, d^2\} \ \log p.$$

  - Logistic regression for discrete models succeeds with sample size:

$$n > c_1 \ \max\{\frac{d}{\theta_{\min}^2}, d^3\} \ \log p.$$

Summary and open questions

- *Practical methods:*
  - Log-determinant for Gaussian graphical models:

$$n > c_1 \ \max\{\frac{1}{\theta_{\min}^2}, d^2\} \ \log p.$$

  - Logistic regression for discrete models succeeds with sample size:

$$n > c_1 \ \max\{\frac{d}{\theta_{\min}^2}, d^3\} \ \log p.$$

- *Fundamental limit:* Any algorithm fails for sample size

$$n < c_2 \ \max\{\frac{1}{\theta_{\min}^2}, d^2\} \ \log p$$

Summary and open questions

- *Practical methods:*
  - Log-determinant for Gaussian graphical models:

$$n > c_1 \ \max\{\frac{1}{\theta_{\min}^2}, d^2\} \ \log p.$$

  - Logistic regression for discrete models succeeds with sample size:

$$n > c_1 \ \max\{\frac{d}{\theta_{\min}^2}, d^3\} \ \log p.$$

- *Fundamental limit:* Any algorithm fails for sample size

$$n < c_2 \ \max\{\frac{1}{\theta_{\min}^2}, d^2\} \ \log p$$

- some extensions and open questions:
  - non-binary discrete MRFs via block-structured regularization schemes
  - other performance metrics (e.g, $(1-\delta)$ edges correct)
  - broader issue: optimal trade-offs between statistical/computational efficiency?

Some papers on graph selection

- Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2009). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*.
- Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. High-dimensional covariance estimation: Convergence rates of ℓ_1-regularized log-determinant divergence. Appeared at *NIPS Conference 2008*.
- Santhanam, P. and Wainwright, M. J. (2008). Information-theoretic limitations of high-dimensional graphical model selection. Presented at *International Symposium on Information Theory,* 2008.
- Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. *IEEE Trans. on Information Theory*, May 2009.

A key separation lemma

Strategy: Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st})\|\theta(G^{uv})) = D\big(\theta(G^{st})\|\theta(G^{uv})\big) + D\big(\theta(G^{uv})\|\theta(G^{st})\big)$$

A key separation lemma

Strategy: Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st}) \| \theta(G^{uv})) = D\big(\theta(G^{st}) \| \theta(G^{uv})\big) + D\big(\theta(G^{uv}) \| \theta(G^{st})\big)$$

Lemma

For the given ensemble, the symmetrized KL divergence is upper bounded as

$$S(\theta(G^{st}) \| \theta(G^{uv})) \leq \frac{8d\theta_{\min} \exp(3\theta_{\min}/2)}{\exp(d\theta_{\min}/2)}$$

A key separation lemma

Strategy: Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st}) \| \theta(G^{uv})) = D\big(\theta(G^{st}) \| \theta(G^{uv})\big) + D\big(\theta(G^{uv}) \| \theta(G^{st})\big)$$

Lemma

For the given ensemble, the symmetrized KL divergence is upper bounded as

$$S(\theta(G^{st}) \| \theta(G^{uv})) \leq \frac{8d\theta_{\min} \exp(3\theta_{\min}/2)}{\exp(d\theta_{\min}/2)}$$

Key consequences:

- complexity controls exponentially in maximum neighborhood weight

$$\omega(\theta^*) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}|.$$

A key separation lemma

Strategy: Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st})\|\theta(G^{uv})) = D\big(\theta(G^{st})\|\theta(G^{uv})\big) + D\big(\theta(G^{uv})\|\theta(G^{st})\big)$$

Lemma

For the given ensemble, the symmetrized KL divergence is upper bounded as

$$S(\theta(G^{st})\|\theta(G^{uv})) \leq \frac{8d\theta_{\min}\exp(3\theta_{\min}/2)}{\exp(d\theta_{\min}/2)}$$

Key consequences:

- complexity controls exponentially in maximum neighborhood weight

$$\omega(\theta^*) := \max_{s\in V} \sum_{t\in N(s)} |\theta_{st}|.$$

- combining with Fano's inequality yields the necessary condition

$$n \; > \; \frac{\exp(\frac{\omega(\theta)}{4})\, d\theta_{\min} \log(pd/8)}{128\exp(\frac{3\theta_{\min}}{2})}$$

Sufficient conditions for $G_{d,p}$

- $G \in \mathcal{G}_{d,p}$: graphs with p nodes and max. degree d
- Ising models with:
  - *Minimum edge weight:* $|\theta^*_{st}| \geq \theta_{\min}$ for all edges
  - *Maximum neighborhood weight:* $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta^*_{st}|$

Sufficient conditions for $G_{d,p}$

- $G \in \mathcal{G}_{d,p}$: graphs with p nodes and max. degree d
- Ising models with:
  - *Minimum edge weight:* $|\theta^*_{st}| \geq \theta_{\min}$ for all edges
  - *Maximum neighborhood weight:* $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta^*_{st}|$

Theorem

There is an (exponential-time) method that succeeds if

$$n \quad > \quad \max\left\{d \log p,\ \frac{6 \exp(2\omega(\theta))}{\sinh^2(\frac{|\theta|}{2})} d \log p,\ \frac{8 \log p}{\theta^2_{\min}}\right\}.$$

Sufficient conditions for $G_{d,p}$

- $G \in \mathcal{G}_{d,p}$: graphs with p nodes and max. degree d
- Ising models with:
  - *Minimum edge weight:* $|\theta^*_{st}| \geq \theta_{\min}$ for all edges
  - *Maximum neighborhood weight:* $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta^*_{st}|$

Theorem

There is an (exponential-time) method that succeeds if

$$n \quad > \quad \max\left\{d \log p,\ \frac{6\exp(2\omega(\theta))}{\sinh^2(\frac{|\theta|}{2})} d \log p,\ \frac{8 \log p}{\theta^2_{\min}}\right\}.$$

Comments:

- to avoid exponential penalty via maximum neighborhood term, require that $\theta_{\min} = \mathcal{O}(1/d)$
- leads to simplified lower bound $n = \Omega\big(\max\{\frac{\log p}{\theta^2_{\min}},\ d^3 \log p\}\big)$