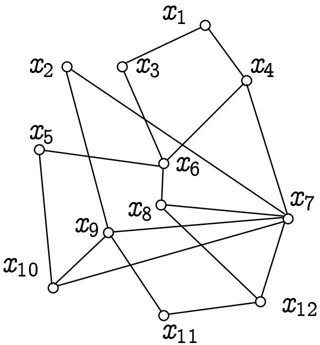
# Stat 375: Inference in Graphical Models
# Lectures 11, 12

Andrea Montanari

Stanford University

May 13, 2012

# Will focus on Ising models




$G = (V, E)$, $V = [n]$, $x = (x_1, \dots, x_n)$, $x_i \in \{+1, -1\}$

$$\mu_{G,\theta}(x) = \frac{1}{Z_{G,\theta}} \exp\Big\{ \sum_{(i,j)\in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \Big\} .$$

# Outline

## Learning graphical models: Setting

# Learning

You are given

$$x^{(1)},\, x^{(2)}, \dots, x^{(n)} \sim_{\text{i.i.d.}} \mu_{G,\theta}(\,\cdot\,)$$

Question: Estimate $\theta$, $G$

# Some notation

Number of vertices $p$

Number of samples $n$

When necessary, 'truth' will be indicated by $\theta^*$

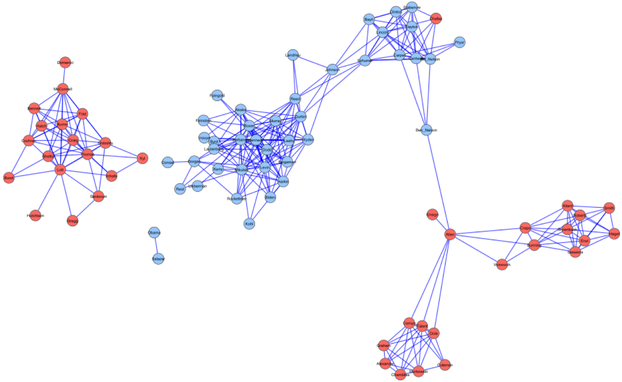
# Three critiques of this setting

- The actual distribution is only *roughly* Ising.
- Variables only observed partially/corrupted.
- Sampling is hard. How can you hope to get i.i.d. samples?

# A toy example: US Senate

| | |
|---|---|
| Number of senators | $p = 100$ |
| Number of bills in 2004 – 2006 | $n = 542$ |
| Voting pattern on bill $p$ | $x^{(\ell)} \in \{+1, -1\}^p$ |

# A toy example: US Senate



[O.Banerjee, L.El Ghaoui, A.d'Aspremont, J.Mach.Learn.Res., 2008]

# Parameter learning

# Is it at all possible?

Define, for each $i \in V$, $\{i,j\} \subseteq V$

$$\begin{aligned}\widetilde{\psi}_i(x_i) &\equiv \frac{\mu(x_i, +1_{V\setminus i})}{\mu(+1_V)}\,,\\ \widetilde{\psi}_{i,j}(x_i,x_j) &\equiv \frac{\mu(x_i,x_j,+1_{V\setminus\{i,j\}})\,\mu(+1_V)}{\mu(x_i,+1_{V\setminus i})\,\mu(x_j,+1_{V\setminus j})}\,.\end{aligned}$$

By Hammersley-Clifford

$$\mu(x) = \mu(+1_V) \prod_{\{i,j\}\in V} \widetilde{\psi}_{ij}(x_i,x_j) \prod_{i\in V} \widetilde{\psi}_i(x_i)\,,$$

$$\widetilde{\psi}_{ij}(\,\cdot\,,\cdot\,) \neq 1 \Rightarrow (i,j) \in E$$

# Is it at all possible?

Define, for each $i \in V$, $\{i,j\} \subseteq V$

$$\begin{aligned}
\widetilde{\psi}_i(x_i) &\equiv \frac{\mu(x_i, +1_{V\setminus i})}{\mu(+1_V)}\,,\\
\widetilde{\psi}_{i,j}(x_i,x_j) &\equiv \frac{\mu(x_i,x_j,+1_{V\setminus\{i,j\}})\,\mu(+1_V)}{\mu(x_i,+1_{V\setminus i})\,\mu(x_j,+1_{V\setminus j})}\,.
\end{aligned}$$

By Hammersley-Clifford

$$\mu(x) = \mu(+1_V)\prod_{\{i,j\}\in V}\widetilde{\psi}_{ij}(x_i,x_j)\prod_{i\in V}\widetilde{\psi}_i(x_i)\,,$$

$$\widetilde{\psi}_{ij}(\,\cdot\,,\cdot\,) \neq 1 \Rightarrow (i,j) \in E$$

# We can estimate these from samples

$$\widehat{\mu}^{(n)}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{I}(x^{(\ell)} = x)$$

$$\widehat{\psi}_i^{(n)}(x_i) \equiv \frac{\widehat{\mu}^{(n)}(x_i, +1_{V\setminus i})}{\widehat{\mu}^{(n)}(+1_V)},$$

$$\widehat{\psi}_{i,j}^{(n)}(x_i, x_j) \equiv \frac{\widehat{\mu}^{(n)}(x_i, x_j, +1_{V\setminus\{i,j\}})\,\widehat{\mu}^{(n)}(+1_V)}{\widehat{\mu}^{(n)}(x_i, +1_{V\setminus i})\,\widehat{\mu}^{(n)}(x_j, +1_{V\setminus j})}.$$

$$\lim_{n\to\infty} \widehat{\psi}^{(n)} = \widetilde{\psi}$$

# We can estimate these from samples

$$\widehat{\mu}^{(n)}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{I}(x^{(\ell)} = x)$$

$$\widehat{\psi}_i^{(n)}(x_i) \equiv \frac{\widehat{\mu}^{(n)}(x_i, +1_{V\setminus i})}{\widehat{\mu}^{(n)}(+1_V)},$$

$$\widehat{\psi}_{i,j}^{(n)}(x_i, x_j) \equiv \frac{\widehat{\mu}^{(n)}(x_i, x_j, +1_{V\setminus\{i,j\}})\, \widehat{\mu}^{(n)}(+1_V)}{\widehat{\mu}^{(n)}(x_i, +1_{V\setminus i})\, \widehat{\mu}^{(n)}(x_j, +1_{V\setminus j})}.$$

$$\lim_{n\to\infty} \widehat{\psi}^{(n)} = \widetilde{\psi}$$

# We can estimate these from samples

$$\widehat{\mu}^{(n)}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{I}(x^{(\ell)} = x)$$

$$\begin{aligned}
\widehat{\psi}_i^{(n)}(x_i) &\equiv \frac{\widehat{\mu}^{(n)}(x_i, +1_{V\setminus i})}{\widehat{\mu}^{(n)}(+1_V)}\,, \\
\widehat{\psi}_{i,j}^{(n)}(x_i, x_j) &\equiv \frac{\widehat{\mu}^{(n)}(x_i, x_j, +1_{V\setminus\{i,j\}})\,\widehat{\mu}^{(n)}(+1_V)}{\widehat{\mu}^{(n)}(x_i, +1_{V\setminus i})\,\widehat{\mu}^{(n)}(x_j, +1_{V\setminus j})}\,.
\end{aligned}$$

$$\lim_{n\to\infty} \widehat{\psi}^{(n)} = \widetilde{\psi}$$

# How big is $\infty$?

**Exercise:** Let $X_1, \dots X_n \sim_{\text{i.i.d.}} \text{Bernoulli}(q)$ with $q \in (0, 1/2]$. The smallest $n$ to estimate $q$ with multiplicative accuracy $\varepsilon$, with probability at least $1 - \delta$ is

$$n \geq \frac{C}{q\varepsilon^2} \log(1/\delta)$$

**Hint:** Use the large deviations estimate

$$\mathbb{P}\Big\{ \frac{1}{n} \sum_{i=1}^{n} X_i \approx r \Big\} \approx e^{-nD(r||q)} \, .$$

# We want this with multiplicative error $\varepsilon$

$$\widehat{\mu}^{(n)}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{I}(x^{(\ell)} = x)$$

$$n \gtrsim \frac{1}{\mu(x)} = e^{\Theta(p)}$$

# We want this with multiplicative error $\varepsilon$

$$\widehat{\mu}^{(n)}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{I}(x^{(\ell)} = x)$$

$$n \gtrsim \frac{1}{\mu(x)} = e^{\Theta(p)}$$

# Wait a minute

For each $i \in V$, $\{i,j\} \subseteq V$

$$\begin{aligned}
\widetilde{\psi}_i(x_i) &\equiv \frac{\mu(x_i, +1_{V\setminus i})}{\mu(+1_V)}\,, \\
\widetilde{\psi}_{i,j}(x_i, x_j) &\equiv \frac{\mu(x_i, x_j, +1_{V\setminus\{i,j\}})\,\mu(+1_V)}{\mu(x_i, +1_{V\setminus i})\,\mu(x_j, +1_{V\setminus j})}\,.
\end{aligned}$$

# Wait a minute

For each $i \in V$, $\{i,j\} \subseteq V$

$$\widetilde{\psi}_i(x_i) \equiv \frac{\mu(x_i, +1_{\partial i})}{\mu(+1_{i\cup\partial i})},$$

$$\widetilde{\psi}_{i,j}(x_i, x_j) \equiv \frac{\mu(x_i, x_j + 1_{\partial ij})\, \mu(+1_{\{i,j\}\cup\partial ij})}{\mu(x_i, +1_{\{j\}\cup\partial ij})\, \mu(x_j, +1_{\{i\}\cup\partial ij})}.$$

Sufficient to estimate $\widehat{\mu}_{\partial ij\cup\{i,j\}}$:

$$n \geq \frac{C}{\varepsilon^2 \min_{|S|\leq 3k, x_S} \mu_S(x_S)} \log(1/\delta)$$

# Wait a minute

For each $i \in V$, $\{i,j\} \subseteq V$

$$\widetilde{\psi}_i(x_i) \equiv \frac{\mu(x_i, +1_{\partial i})}{\mu(+1_{i \cup \partial i})},$$

$$\widetilde{\psi}_{i,j}(x_i, x_j) \equiv \frac{\mu(x_i, x_j + 1_{\partial ij})\, \mu(+1_{\{i,j\} \cup \partial ij})}{\mu(x_i, +1_{\{j\} \cup \partial ij})\, \mu(x_j, +1_{\{i\} \cup \partial ij})}.$$

Sufficient to estimate $\widehat{\mu}_{\partial ij \cup \{i,j\}}$:

$$n \geq \frac{C}{\varepsilon^2 \min_{|S| \leq 3k, x_S} \mu_S(x_S)} \log(1/\delta)$$

# Sample complexity

**Theorem** (P.Abeel, D.Koller, A.Ng, J. Mach. Learn. Res., 2006)

*Assume*

- $\max_{i\in V} deg_G(i) = k$;
- $\min_{(ij)\in E} \min_{x_i,x_j} \psi_{ij}(x_i,x_j) \geq \rho$, $\min_{i\in V} \min_{x_i} \psi_i(x_i) \geq \rho$.

*Then, with probability at least $1-\delta$, we can learn all the $\psi$'s with relative accuracy $\varepsilon$ provided*

$$n \geq \frac{C(\rho)^k}{\varepsilon^2} \log\left(\frac{p}{\delta}\right).$$

*Further, under these assumptions $D(\mu||\widehat{\mu}) + D(\widehat{\mu}||\mu) \leq C\,|E|\varepsilon$.*

**Proof.**

Lower bound $\min_{|S|\leq 3k, x_S} \mu_S(x_S)$ plus union bound. □

# Sample complexity

**Theorem** (P.Abeel, D.Koller, A.Ng, J. Mach. Learn. Res., 2006)

*Assume*

- $\max_{i \in V} deg_G(i) = k$;
- $\min_{(ij)\in E} \min_{x_i,x_j} \psi_{ij}(x_i,x_j) \geq \rho$, $\min_{i\in V} \min_{x_i} \psi_i(x_i) \geq \rho$.

*Then, with probability at least $1-\delta$, we can learn all the $\psi$'s with relative accuracy $\varepsilon$ provided*

$$n \geq \frac{C(\rho)^k}{\varepsilon^2} \log\left(\frac{p}{\delta}\right).$$

*Further, under these assumptions $D(\mu||\widehat{\mu}) + D(\widehat{\mu}||\mu) \leq C\,|E|\varepsilon$.*

**Proof.**

Lower bound $\min_{|S|\leq 3k, x_S} \mu_S(x_S)$ plus union bound. □

# How good is this?

- Number of bits to specify the $\psi$'s $= |E| \log(1/(\rho\varepsilon))$.
- Number of bits per sample $= p$

$$n \geq \frac{|E|}{p} \log\left(\frac{1}{\rho\varepsilon}\right) = \frac{k}{2} \log\left(\frac{1}{\rho\varepsilon}\right)$$

# How good is this?

- Number of bits to specify the $\psi$'s $= |E| \log(1/(\rho\varepsilon))$.

- Number of bits per sample $= p$

$$n \geq \frac{|E|}{p} \log\left(\frac{1}{\rho\varepsilon}\right) = \frac{k}{2} \log\left(\frac{1}{\rho\varepsilon}\right)$$

# Pretty good, isn't it?

# Pretty good, isn't it?

We are assuming that $G$ is known!

[But see later]

# Maximum likelihood

# Another approach

Likelihood

$$\mathbb{P}_{n,G,\theta}\{x^{(1)},\dots,x^{(n)}\} = \prod_{\ell=1}^{n} \mu_{G,\theta}(x^{(\ell)})$$

$$= \frac{1}{Z_G(\theta)^n} \exp\Big\{ \sum_{(i,j)\in E} \theta_{ij} \sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)} + \sum_{i\in V} \theta_i \sum_{\ell=1}^{n} x_i^{(\ell)} \Big\}$$

$$\mathcal{L}_n(\theta;\{x^{(\ell)}\}) \equiv -\frac{1}{n}\log \mathbb{P}_{n,G,\theta}\{x^{(1)},\dots,x^{(n)}\}$$

$$= -\langle \widehat{M}, \theta\rangle + \phi(\theta)$$

$$\widehat{M}_i = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)}, \qquad \widehat{M}_{ij} = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}$$

# Another approach

Likelihood

$$\mathbb{P}_{n,G,\theta}\{x^{(1)},\dots,x^{(n)}\} = \prod_{\ell=1}^{n} \mu_{G,\theta}(x^{(\ell)})$$

$$= \frac{1}{Z_G(\theta)^n} \exp\Big\{ \sum_{(i,j)\in E} \theta_{ij} \sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)} + \sum_{i\in V} \theta_i \sum_{\ell=1}^{n} x_i^{(\ell)} \Big\}$$

$$\begin{aligned}
\mathcal{L}_n(\theta;\{x^{(\ell)}\}) &\equiv -\frac{1}{n}\log \mathbb{P}_{n,G,\theta}\{x^{(1)},\dots,x^{(n)}\} \\
&= -\langle \widehat{M}, \theta\rangle + \phi(\theta) \\
&\quad \widehat{M}_i = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)}, \qquad \widehat{M}_{ij} = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}
\end{aligned}$$

# Maximum likelihood

$$\widehat{\theta}^{(n)} = \widehat{\theta}(\{x^{(\ell)}\}_{1\le\ell\le n}) \equiv \arg\min_{\theta} \mathcal{L}_n(\theta;\{x^{(\ell)}\})$$

*Unique* : $\theta \mapsto \mathcal{L}_n(\theta;\{x^{(\ell)}\})$ is strictly convex;

*Consistent* : As $n \to \infty$, $\widehat{M} \to M$, $M_i = \mathbb{E}_{\theta_*}\{x_i\}$, $M_{ij} = \mathbb{E}_{\theta_*}\{x_i x_j\}$

# Maximum likelihood

$$\widehat{\theta}^{(n)} = \widehat{\theta}(\{x^{(\ell)}\}_{1\le\ell\le n}) \equiv \arg\min_{\theta} \mathcal{L}_n(\theta; \{x^{(\ell)}\})$$

*Unique* : $\theta \mapsto \mathcal{L}_n(\theta; \{x^{(\ell)}\})$ is strictly convex;

*Consistent* : As $n \to \infty$, $\widehat{M} \to M$, $M_i = \mathbb{E}_{\theta_*}\{x_i\}$, $M_{ij} = \mathbb{E}_{\theta_*}\{x_i x_j\}$

# Maximum likelihood

$$\widehat{\theta}^{(n)} = \widehat{\theta}(\{x^{(\ell)}\}_{1\le\ell\le n}) \equiv \arg\min_{\theta} \mathcal{L}_n(\theta; \{x^{(\ell)}\})$$

*Unique* : $\theta \mapsto \mathcal{L}_n(\theta; \{x^{(\ell)}\})$ is strictly convex;

*Consistent* : As $n \to \infty$, $\widehat{M} \to M$, $M_i = \mathbb{E}_{\theta_*}\{x_i\}$, $M_{ij} = \mathbb{E}_{\theta_*}\{x_i x_j\}$

# Proof of consistency

$$\phi(\widehat{\theta}) - \langle \widehat{M}, \widehat{\theta} \rangle \leq \phi(\theta) - \langle \widehat{M}, \theta \rangle$$

By strict convexity, for $\xi > 0$

$$\phi(\theta') \geq \phi(\theta) + \langle M, (\theta' - \theta) \rangle + \frac{\xi}{2} \|\theta' - \theta\|_2^2$$

Hence

$$\langle M, (\widehat{\theta} - \theta) \rangle + \frac{\xi}{2} \|\widehat{\theta} - \theta\|_2^2 - \langle \widehat{M}, \widehat{\theta} \rangle \leq -\langle \widehat{M}, \theta \rangle$$

$$\frac{\xi}{2} \|\widehat{\theta} - \theta\|_2^2 \leq \langle (\widehat{M} - M), (\widehat{\theta} - \theta) \rangle \leq \|\widehat{M} - M\|_2 \|\widehat{\theta} - \theta\|_2$$

# Proof of consistency

$$\phi(\widehat{\theta}) - \langle \widehat{M}, \widehat{\theta} \rangle \leq \phi(\theta) - \langle \widehat{M}, \theta \rangle$$

By strict convexity, for $\xi > 0$

$$\phi(\theta') \geq \phi(\theta) + \langle M, (\theta' - \theta) \rangle + \frac{\xi}{2} ||\theta' - \theta||_2^2$$

Hence

$$\langle M, (\widehat{\theta} - \theta) \rangle + \frac{\xi}{2} ||\widehat{\theta} - \theta||_2^2 - \langle \widehat{M}, \widehat{\theta} \rangle \leq -\langle \widehat{M}, \theta \rangle$$

$$\frac{\xi}{2} ||\widehat{\theta} - \theta||_2^2 \leq \langle (\widehat{M} - M), (\widehat{\theta} - \theta) \rangle \leq ||\widehat{M} - M||_2 ||\widehat{\theta} - \theta||_2$$

# Proof of consistency

$$\phi(\widehat{\theta}) - \langle \widehat{M}, \widehat{\theta} \rangle \le \phi(\theta) - \langle \widehat{M}, \theta \rangle$$

By strict convexity, for $\xi > 0$

$$\phi(\theta') \ge \phi(\theta) + \langle M, (\theta' - \theta) \rangle + \frac{\xi}{2} \|\theta' - \theta\|_2^2$$

Hence

$$\langle M, (\widehat{\theta} - \theta) \rangle + \frac{\xi}{2} \|\widehat{\theta} - \theta\|_2^2 - \langle \widehat{M}, \widehat{\theta} \rangle \le -\langle \widehat{M}, \theta \rangle$$

$$\frac{\xi}{2} \|\widehat{\theta} - \theta\|_2^2 \le \langle (\widehat{M} - M), (\widehat{\theta} - \theta) \rangle \le \|\widehat{M} - M\|_2 \|\widehat{\theta} - \theta\|_2$$

# Proof of consistency

$$\phi(\widehat{\theta}) - \langle \widehat{M}, \widehat{\theta} \rangle \leq \phi(\theta) - \langle \widehat{M}, \theta \rangle$$

By strict convexity, for $\xi > 0$

$$\phi(\theta') \geq \phi(\theta) + \langle M, (\theta' - \theta) \rangle + \frac{\xi}{2} ||\theta' - \theta||_2^2$$

Hence

$$\langle M, (\widehat{\theta} - \theta) \rangle + \frac{\xi}{2} ||\widehat{\theta} - \theta||_2^2 - \langle \widehat{M}, \widehat{\theta} \rangle \leq -\langle \widehat{M}, \theta \rangle$$

$$\frac{\xi}{2} ||\widehat{\theta} - \theta||_2^2 \leq \langle (\widehat{M} - M), (\widehat{\theta} - \theta) \rangle \leq ||\widehat{M} - M||_2 ||\widehat{\theta} - \theta||_2$$

# High-dimensional consistency bound

$|E| = m$

$$\frac{1}{m}\|\widehat{\theta} - \theta\|_2^2 \leq \frac{4}{\xi^2 m}\|\widehat{M} - M\|_2^2$$

$$\leq \frac{C}{\xi^2\, n\, m}\binom{p}{2}$$

$$\xi \equiv \inf_{\theta} \sigma_{\min}(\nabla^2 \phi(\theta))$$

with more work can eliminate the $\inf_\theta$.

# High-dimensional consistency bound

$|E| = m$

$$\frac{1}{m}\|\widehat{\theta} - \theta\|_2^2 \le \frac{4}{\xi^2 m}\|\widehat{M} - M\|_2^2 \le \frac{C}{\xi^2\, n\, m}\binom{p}{2}$$

$$\xi \equiv \inf_{\theta} \sigma_{\min}(\nabla^2 \phi(\theta))$$

with more work can eliminate the $\inf_\theta$.

# High-dimensional consistency bound

$|E| = m$

$$\frac{1}{m}\|\widehat{\theta} - \theta\|_2^2 \le \frac{4}{\xi^2 m}\|\widehat{M} - M\|_2^2 \le \frac{C}{\xi^2\, n\, m}\binom{p}{2}$$

$$\xi \equiv \inf_{\theta} \sigma_{\min}(\nabla^2 \phi(\theta))$$

with more work can eliminate the $\inf_\theta$.

# Sounds good, right?

# Sounds good, right?

Approximating $\phi(\theta)$ is hard!

Bad sample complexity for sparse graphs!

## Structural learning

# Structural learning

$$n_{\mathsf{Alg}}(G,\theta) \equiv \inf\left\{n \in \mathbb{N} : \mathbb{P}_{n,G,\theta}\{\mathsf{Alg}(x^{(1)},\dots,x^{(n)}) = G\} \geq 1-\delta\right\},$$

$$\chi_{\mathsf{Alg}}(G,\theta) \equiv \#\text{ operations of Alg when run on } n_{\mathsf{Alg}}(G,\theta) \text{ samples}$$

Typically, we assume $G$ sparse

# How would you modify maximum likelihod?

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}(\theta; \{x^{(\ell)}\}) \\ \text{subject to} \quad & ||\theta||_0 \le m \end{aligned}$$

Intractable!

# How would you modify maximum likelihod?

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}(\theta; \{x^{(\ell)}\}) \\ \text{subject to} \quad & ||\theta||_0 \le m \end{aligned}$$

Intractable!

# $\ell_1$-regularized maximum likelihood

$$
\begin{aligned}
\widehat{\theta} &= \arg\min_{\theta} \; \mathcal{L}(\theta; \{x^{(\ell)}) + \lambda \|\theta\|_1 \\
&= -\langle \widehat{M}, \theta \rangle + \phi(\theta) + \lambda \|\theta\|_1
\end{aligned}
$$

[cf. J.Friedman, T.Hastie, R.Tibshirani, Biostatistics, 2008]

# $\ell_1$-regularized maximum likelihood

$$\begin{aligned}\widehat{\theta} &= \arg\min_{\theta} \; \mathcal{L}(\theta; \{x^{(\ell)}) + \lambda\|\theta\|_1 \\ &= -\langle \widehat{M}, \theta\rangle + \phi(\theta) + \lambda\|\theta\|_1\end{aligned}$$

[cf. J.Friedman, T.Hastie, R.Tibshirani, Biostatistics, 2008]

# $\ell_1$-regularized maximum likelihood

$$
\begin{aligned}
\widehat{\theta} &= \arg\min_{\theta} \; \mathcal{L}(\theta; \{x^{(\ell)}) + \lambda \|\theta\|_1 \\
&= -\langle \widehat{M}, \theta \rangle + \phi(\theta) + \lambda \|\theta\|_1
\end{aligned}
$$

[cf. J.Friedman, T.Hastie, R.Tibshirani, Biostatistics, 2008]

# Local independence test

Idea: For each $i \in V$, and for any candidate neighborood $S$,

test independence of $x_i$ and $x_{V\setminus S_i}$, $S_i \equiv S \cup \{i\}$.

# A possible implementation

```
LOCAL INDEPENDENCE TEST( samples {x^(ℓ)} )
For each i ∈ V;
    For each S ⊆ V \ {i}, |S| ≤ k,
        Compute SCORE(S, i) = Ĥ(X_i|X_S);
    Set S* = arg min_S SCORE(S, i) and connect i to all j ∈ S*;
Prune the resulting graph.
```

[P.Abeel, D.Koller, A.Ng, 2006]

# Another implementation

$$\textsc{Score}(S,i) \equiv \min_{W\subseteq V\setminus S, j\in S} \max_{x_i,x_W,x_S,x_j} \Big|\widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i | X_W = x_W, X_S = x_S\} - \widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i | X_W = x_W, X_{S\setminus j} = x_{S\setminus j}, X_j = z_j\}\Big| .$$

[G.Bresler, E.Mossel and A.Sly, APPROX 2008]

# Another implementation

| LOCAL INDEPENDENCE TEST( samples $\{x^{(\ell)}\}$, thresholds $(\varepsilon, \gamma)$ ) |
|---|
| 1: For each $i \in V$; |
| 2: For each $S \subseteq V \setminus \{i\}$, $\|S\| \le k$, |
| 3: Compute SCORE$(S, i)$; |
| 4: $S^* = \arg\max\{\|S\| : \text{SCORE}(S, i) > \varepsilon\}$ and connect $i$ to all $j \in S^*$; |

Nobody would ever use these in practice!!

# Another implementation

| Local Independence Test( samples $\{x^{(\ell)}\}$, thresholds $(\varepsilon, \gamma)$ ) |
|---|
| 1: For each $i \in V$; |
| 2: For each $S \subseteq V \setminus \{i\}$, $\vert S\vert \le k$, |
| 3: Compute Score$(S, i)$; |
| 4: $S^* = \arg\max\{\vert S\vert : \text{Score}(S,i) > \varepsilon\}$ and connect $i$ to all $j \in S^*$; |

Nobody would ever use these in practice!!

# Why?

$n^{k+1}$ operations!

# For the sake of simplicity

$\theta_{ij} = \beta$, $\theta_i = 0$

$$\mu_{G,\beta}(x) = \frac{1}{Z_G(\beta)} \exp\Big\{\beta \sum_{(i,j)\in E} x_i x_j\Big\}$$

$$M = (M_{ij})_{1\le i,j\le n}\,, \quad M_{ij} = \mathbb{E}_{G,\beta}\{x_i x_j\}\,, \quad \widehat{M}_{ij} = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}$$

# For the sake of simplicity

$\theta_{ij} = \beta$, $\theta_i = 0$

$$\mu_{G,\beta}(x) = \frac{1}{Z_G(\beta)} \exp\Big\{\beta \sum_{(i,j)\in E} x_i x_j\Big\}$$

$$M = (M_{ij})_{1\le i,j\le n}\,, \quad M_{ij} = \mathbb{E}_{G,\beta}\{x_i x_j\}\,, \quad \widehat{M}_{ij} = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}$$

# A very simple algorithm

---

THRESHOLDING( samples $\{x^{(\ell)}\}$, threshold $\tau$ )

---

1: Compute the empirical correlations $\{\widehat{M}_{ij}\}_{(i,j)\in V\times V}$;
2: For each $(i,j) \in V \times V$
3: If $\widehat{M}_{ij} \geq \tau$, set $(i,j) \in E$;

# And its analysis

**Theorem**

*If $G$ is a tree, and $\tau(\beta) = (\tanh\beta + \tanh^2\beta)/2$, then*

$$n_{\mathsf{Thr}(\tau)}(G,\theta) \le \frac{8}{(\tanh\beta - \tanh^2\beta)^2}\,\log\frac{2p}{\delta}\,.$$

**Theorem**

*If $G$ has maximum degree $k > 1$ and if $\beta < \mathrm{atanh}(1/(2k))$ then*

$$n_{\mathsf{Thr}(\tau)}(G,\beta) \le \frac{8}{(\tanh\beta - \frac{1}{2k})^2}\,\log\frac{2p}{\delta}\,.$$

**Theorem**

*If $k > 3$ and $\theta > C/k$, there are graphs such that for any $\tau$, $n_{\mathsf{Thr}(\tau)} = \infty$.*

# And its analysis

**Theorem**

*If $G$ is a tree, and $\tau(\beta) = (\tanh\beta + \tanh^2\beta)/2$, then*

$$n_{\mathsf{Thr}(\tau)}(G,\theta) \le \frac{8}{(\tanh\beta - \tanh^2\beta)^2}\,\log\frac{2p}{\delta}\,.$$

**Theorem**

*If $G$ has maximum degree $k > 1$ and if $\beta < \operatorname{atanh}(1/(2k))$ then*

$$n_{\mathsf{Thr}(\tau)}(G,\beta) \le \frac{8}{(\tanh\beta - \frac{1}{2k})^2}\,\log\frac{2p}{\delta}\,.$$

**Theorem**

*If $k > 3$ and $\theta > C/k$, there are graphs such that for any $\tau$, $n_{\mathsf{Thr}(\tau)} = \infty$.*

# And its analysis

## Theorem

*If $G$ is a tree, and $\tau(\beta) = (\tanh\beta + \tanh^2\beta)/2$, then*

$$n_{\mathsf{Thr}(\tau)}(G, \theta) \le \frac{8}{(\tanh\beta - \tanh^2\beta)^2} \, \log\frac{2p}{\delta} \, .$$

## Theorem

*If $G$ has maximum degree $k > 1$ and if $\beta < \operatorname{atanh}(1/(2k))$ then*

$$n_{\mathsf{Thr}(\tau)}(G, \beta) \le \frac{8}{(\tanh\beta - \frac{1}{2k})^2} \, \log\frac{2p}{\delta} \, .$$

## Theorem

*If $k > 3$ and $\theta > C/k$, there are graphs such that for any $\tau$, $n_{\mathsf{Thr}(\tau)} = \infty$.*

# Basic intuition

Thresholding works if

$$\min_{(i,j)\in E} M_{ij} > \max_{(kl)\notin E} M_{kl}$$

This is true at small $\beta$ because...

# High temperature series

$$Z_G(\beta) = \sum_{H \subseteq G, \text{even}} \tau^{|E(H)|},$$

$$\mathbb{E}_{G,\beta}\{x_i x_j\} = \frac{1}{Z_G(\beta)} \sum_{H \subseteq G, \text{odd}(H)=\{i,j\},} \tau^{|E(H)|},$$

$$\tau = \tanh \beta\,.$$

**Theorem** (R.Griffiths, J. Math. Phys., 1967)

$$\mathbb{E}_{G,\beta}\{x_i x_j\} \le \sum_{\gamma \in \text{SAW}(i \to j)} \tau^{|\gamma|}$$

# High temperature series

$$Z_G(\beta) = \sum_{H\subseteq G,\text{even}} \tau^{|E(H)|}\,,$$

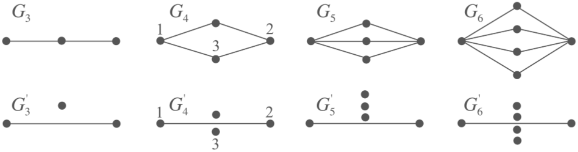
$$\mathbb{E}_{G,\beta}\{x_i x_j\} = \frac{1}{Z_G(\beta)} \sum_{H\subseteq G,\text{odd}(H)=\{i,j\},} \tau^{|E(H)|}\,,$$

$$\tau = \tanh\beta\,.$$

**Theorem** (R.Griffiths, J. Math. Phys., 1967)

$$\mathbb{E}_{G,\beta}\{x_i x_j\} \le \sum_{\gamma\in\text{SAW}(i\to j)} \tau^{|\gamma|}$$

# Does not work always because



# This phenomenon is generic

*Example:* Regularized pseudo-likelihoods
[Meinshausen , Bühlmann, Ann.Stat. 2006]
[P.Ravikumar, M.Wainwright, J.Lafferty, Ann.Stat. 2010]

$\theta_{(i)} \equiv \{\theta_{i,j} : j \in [p] \setminus \{i\}\}$.

$$\text{minimize} \qquad -\frac{1}{n}\sum_{\ell=1}^{n} \log \mathbb{P}_\theta\{x_i^{(\ell)}|x_{\partial i}^{(\ell)}\} + \lambda\|\theta_{(i)}\|_1$$

The first therm only depends on $\theta_{(i)}$! Has explicit expression!

# This phenomenon is generic

*Example:* Regularized pseudo-likelihoods
[Meinshausen , Bühlmann, Ann.Stat. 2006]
[P.Ravikumar, M.Wainwright, J.Lafferty, Ann.Stat. 2010]

$\theta_{(i)} \equiv \{\theta_{i,j} : j \in [p] \setminus \{i\}\}.$

$$\text{minimize} \qquad -\frac{1}{n}\sum_{\ell=1}^{n} \log \mathbb{P}_\theta\{x_i^{(\ell)} | x_{\partial i}^{(\ell)}\} + \lambda \|\theta_{(i)}\|_1$$

The first therm only depends on $\theta_{(i)}$! Has explicit expression!

# This phenomenon is generic

*Example:* Regularized pseudo-likelihoods
[Meinshausen , Bühlmann, Ann.Stat. 2006]
[P.Ravikumar, M.Wainwright, J.Lafferty, Ann.Stat. 2010]

$\theta_{(i)} \equiv \{\theta_{i,j} : j \in [p] \setminus \{i\}\}.$

$$\text{minimize} \qquad -\frac{1}{n}\sum_{\ell=1}^{n} \log \mathbb{P}_\theta\{x_i^{(\ell)} | x_{\partial i}^{(\ell)}\} + \lambda \|\theta_{(i)}\|_1$$

The first therm only depends on $\theta_{(i)}$! Has explicit expression!

# This phenomenon is generic

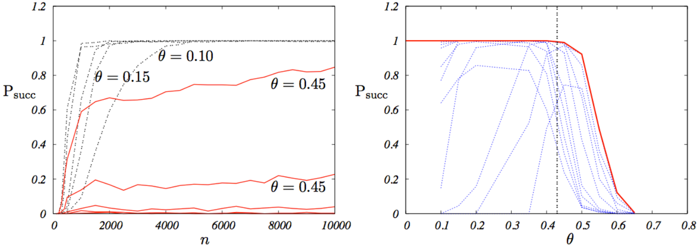
*Example:* Regularized pseudo-likelihoods
[Meinshausen , Bühlmann, Ann.Stat. 2006]
[P.Ravikumar, M.Wainwright, J.Lafferty, Ann.Stat. 2010]

$\theta_{(i)} \equiv \{\theta_{i,j} : j \in [p] \setminus \{i\}\}$.

$$\text{minimize} \qquad -\frac{1}{n}\sum_{\ell=1}^{n} \log \mathbb{P}_\theta\{x_i^{(\ell)}|x_{\partial i}^{(\ell)}\} + \lambda||\theta_{(i)}||_1$$

The first therm only depends on $\theta_{(i)}$! Has explicit expression!

# A numerical experiment




Uniformly random graphs of degree $k = 4$

# A numerical experiment

**Theorem** (J.Bento, A.Montanari, 2010)

*There exists $C > 0$ such that regularized pseudolikelihod fails*

- *On random regular graphs of deree $k$ for all $\beta > C/k$.*
- *On random subgraphs of the 2d grid, for all $\beta > C$.*
- *On 'double-star' graphs for all $\beta > C$, and $n > n_0$*

# Structural learning

- Lots of open problems