Selection for discrete graphical models: Generalized covariance matrices and their inverses

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Based on joint work with:

Po-Ling Loh (UC Berkeley)

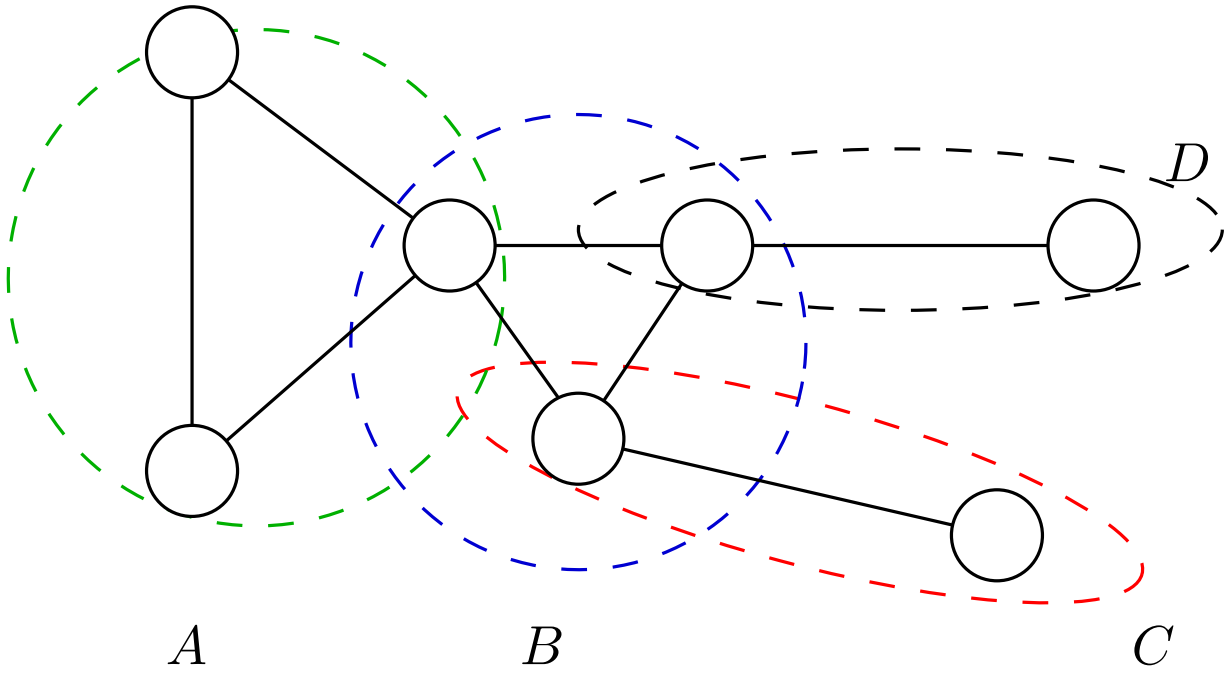
Graphical models for high-dimensional data

Used in a variety of applications:

- bioinformatics, genetics, proteonomics
- natural language processing, speech processing
- social networks
- image processing, computer vision
- satisfiability problems, channel communications, data compression

Graph structure and factorization

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



- Hammersley-Clifford theorem: factorization over cliques

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

Fitting graphical models to data

- drawn n samples from

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

- graph $G = (V, E)$ and clique potentials $\{\theta_C\}$ are unknown

Fitting graphical models to data

- drawn n samples from

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

- graph $G = (V, E)$ and clique potentials $\{\theta_C\}$ are unknown

- use n samples to form data matrix $\mathbf{X} \in \mathcal{X}^{n \times p}$

- estimator $\mathbf{X} \mapsto (\widehat{E}, \{\widehat{\theta}_C\})$

Fitting graphical models to data

- drawn n samples from

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

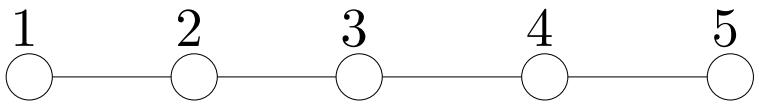
- graph $G = (V, E)$ and clique potentials $\{\theta_C\}$ are unknown

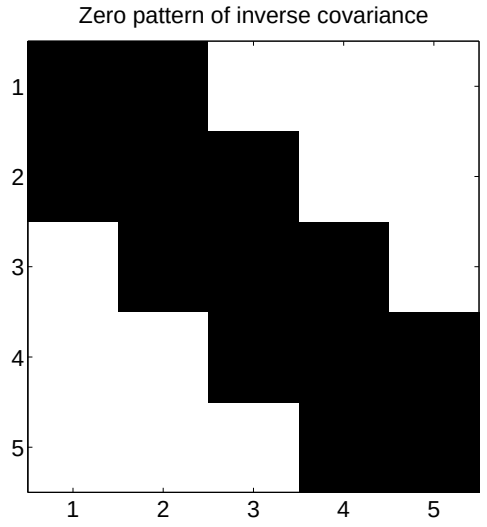
- use n samples to form data matrix $\mathbf{X} \in \mathcal{X}^{n \times p}$

- estimator $\mathbf{X} \mapsto (\widehat{E}, \{\widehat{\theta}_C\})$

- various loss functions are possible:
  - exact graph selection: $\widehat{E} = E$?
  - Hamming distance between $\widehat{E}$ and E
  - parameter estimation $\|\widehat{\theta} - \theta\|$
  - bounds on Kullback-Leibler divergence $D(\mathbb{Q}_{\widehat{\theta}} \,\|\, \mathbb{Q}_\theta)$

Illustration: Graphs and Gaussian inverse covariance





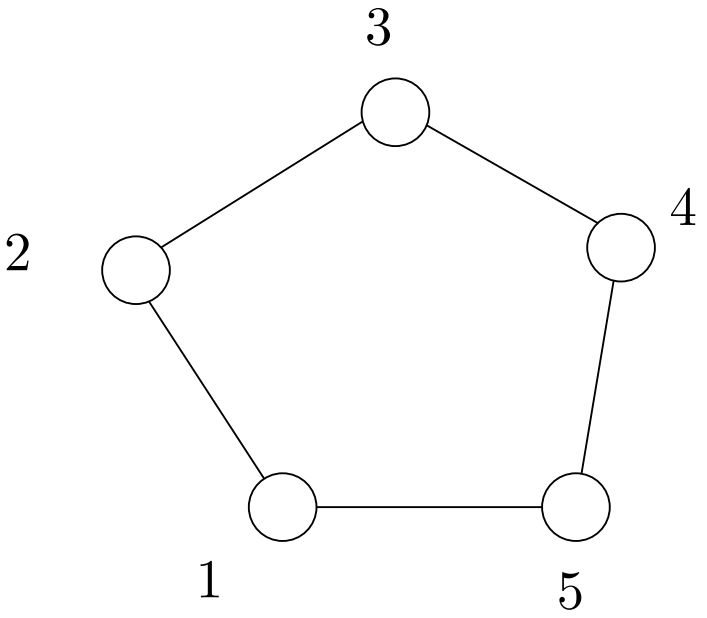
Multivariate Gaussian $(X_1, \ldots, X_p) \sim N(0, \Theta^{-1})$:

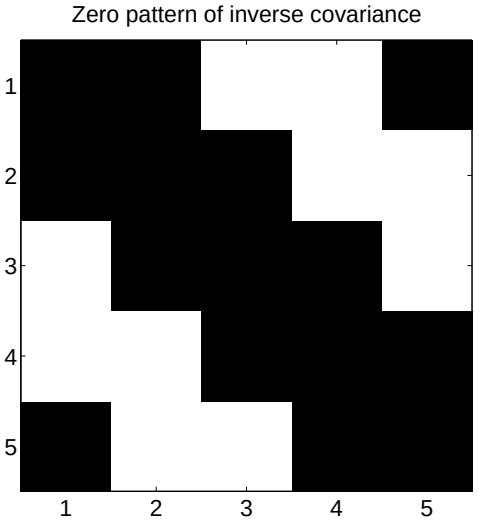
$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\Big(- \sum_{s \in V} \theta_{ss} x_s^2 - \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big).$$

Classical fact:

Inverse covariance matrix Θ of any multivariate Gaussian is graph-structured. (Consequence of Hammersley-Clifford theorem).

Illustration: Graphs and Gaussian inverse covariance





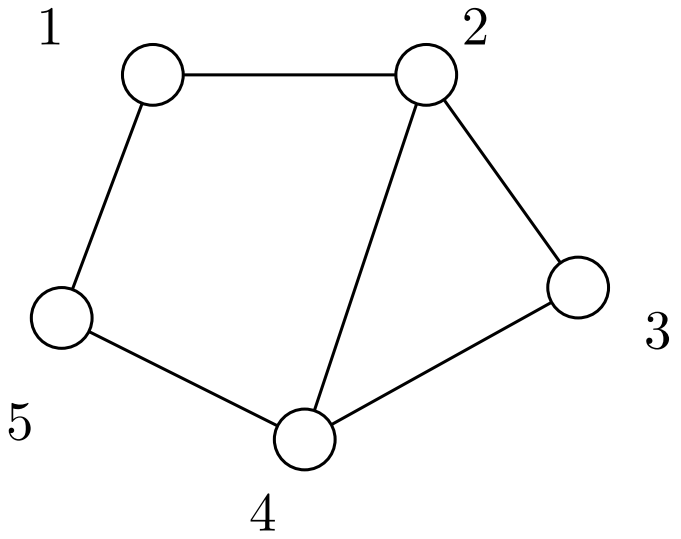
Multivariate Gaussian $(X_1, \ldots, X_p) \sim N(0, \Theta^{-1})$:

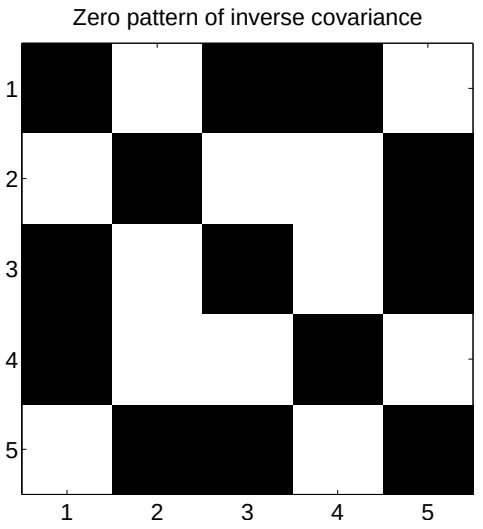
$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\Big(- \sum_{s \in V} \theta_{ss} x_s^2 - \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big).$$

Classical fact:

Inverse covariance matrix Θ of any multivariate Gaussian is graph-structured. (Consequence of Hammersley-Clifford theorem).

Illustration: Graphs and Gaussian inverse covariance





Multivariate Gaussian $(X_1, \ldots, X_p) \sim N(0, \Theta^{-1})$:

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\Big(- \sum_{s \in V} \theta_{ss} x_s^2 - \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big).$$

Classical fact:

Inverse covariance matrix Θ of any multivariate Gaussian is graph-structured. (Consequence of Hammersley-Clifford theorem).

Utility for Gaussian models

Sparsity of inverse covariance can be exploited in many ways:

- global Gaussian likelihood (log-determinant) with ℓ_1-regularization (d'Aspremont et al., 2007; Friedman et al., 2008; Ravikumar et al., 2009)
- various neighborhood-based approaches:
  - neighborhood Lasso method (Meinshausen & Buhlmann, 2006)
  - neighborhood Dantzig selector (Yuan, 2010)
- CLIME estimator for inverse covariance Θ (Cai et al., 2011)
- non-convex Lasso for Gaussians with missing data (Loh & Wainwright, 2011)
- copula methods: univariate transformations of Gaussians (Liu, Lafferty & Wasserman, 2011)

Utility for Gaussian models

Sparsity of inverse covariance can be exploited in many ways:

- global Gaussian likelihood (log-determinant) with ℓ_1-regularization (d'Aspremont et al., 2007; Friedman et al., 2008; Ravikumar et al., 2009)
- various neighborhood-based approaches:
  - neighborhood Lasso method (Meinshausen & Buhlmann, 2006)
  - neighborhood Dantzig selector (Yuan, 2010)
- CLIME estimator for inverse covariance Θ (Cai et al., 2011)
- non-convex Lasso for Gaussians with missing data (Loh & Wainwright, 2011)
- copula methods: univariate transformations of Gaussians (Liu, Lafferty & Wasserman, 2011)

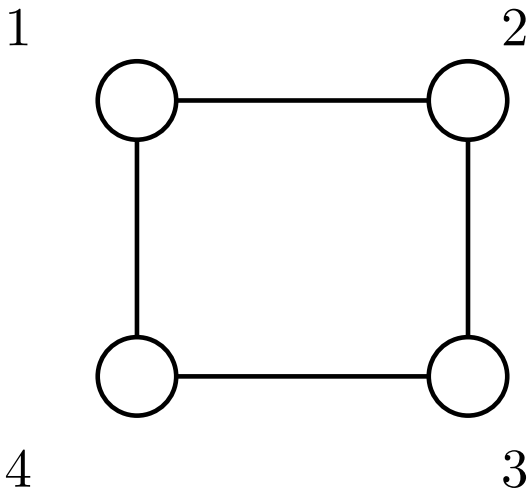
A natural question:

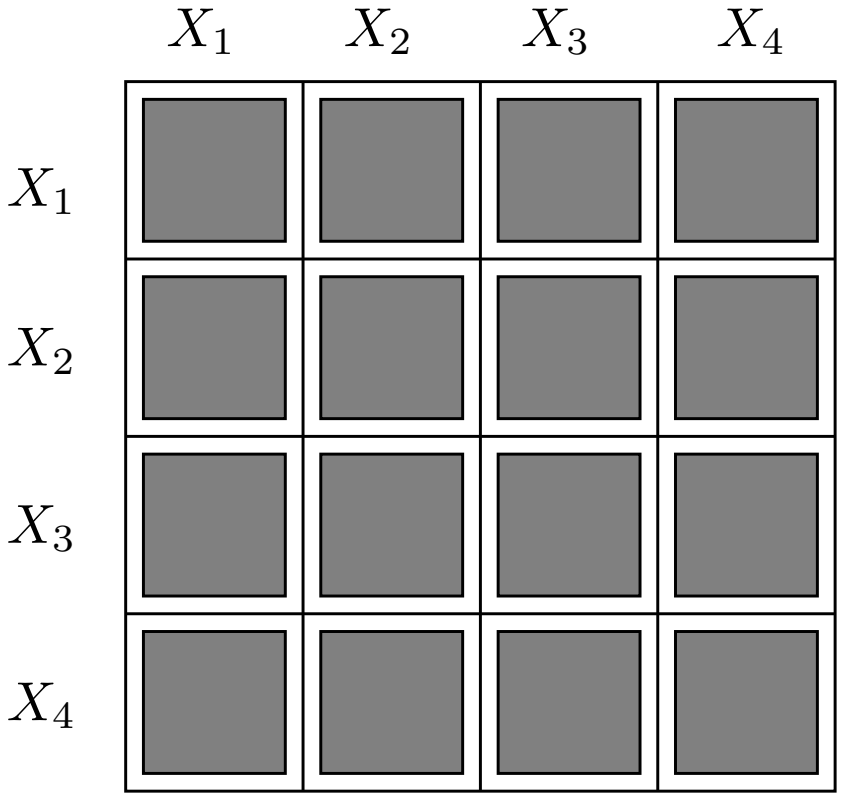
Is there any analog of this type of correspondence for non-Gaussian models?

Simplest version is for the *Ising model*:

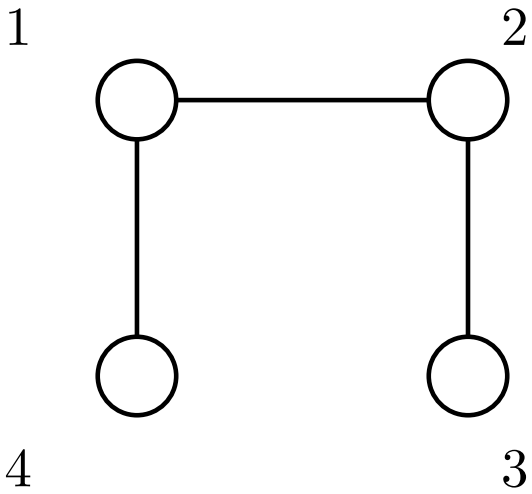
$$\mathbb{Q}_\theta(x_1, \ldots, x_p) \propto \exp\Big\{ \sum \theta_s x_s + \sum \theta_{st} x_s x_t \Big\} \qquad \text{where } X_s \in \{-1, +1\}.$$

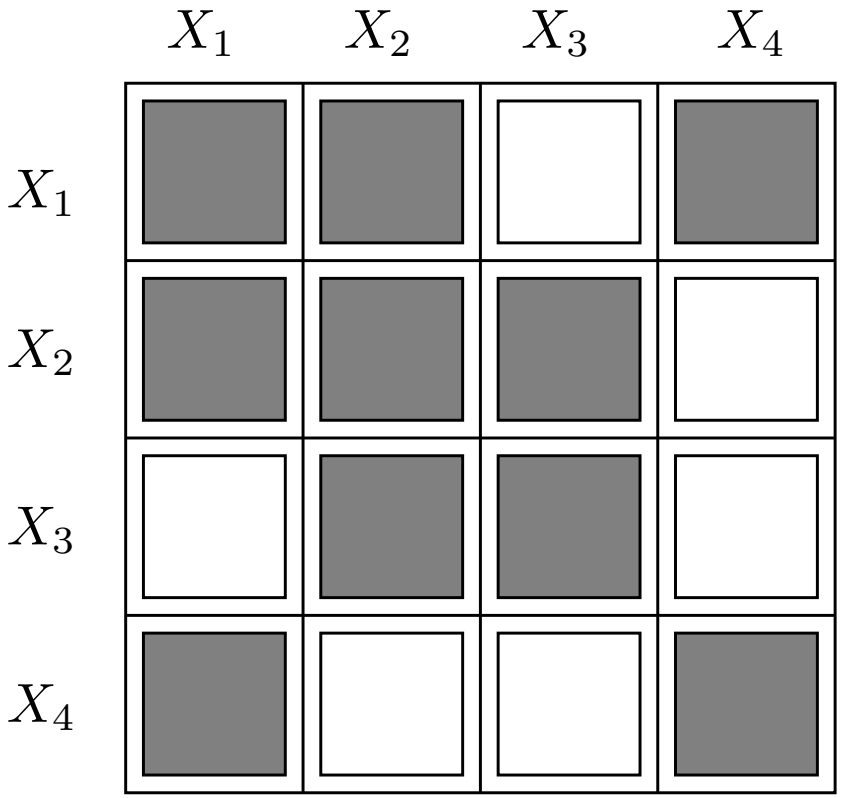
A counter example



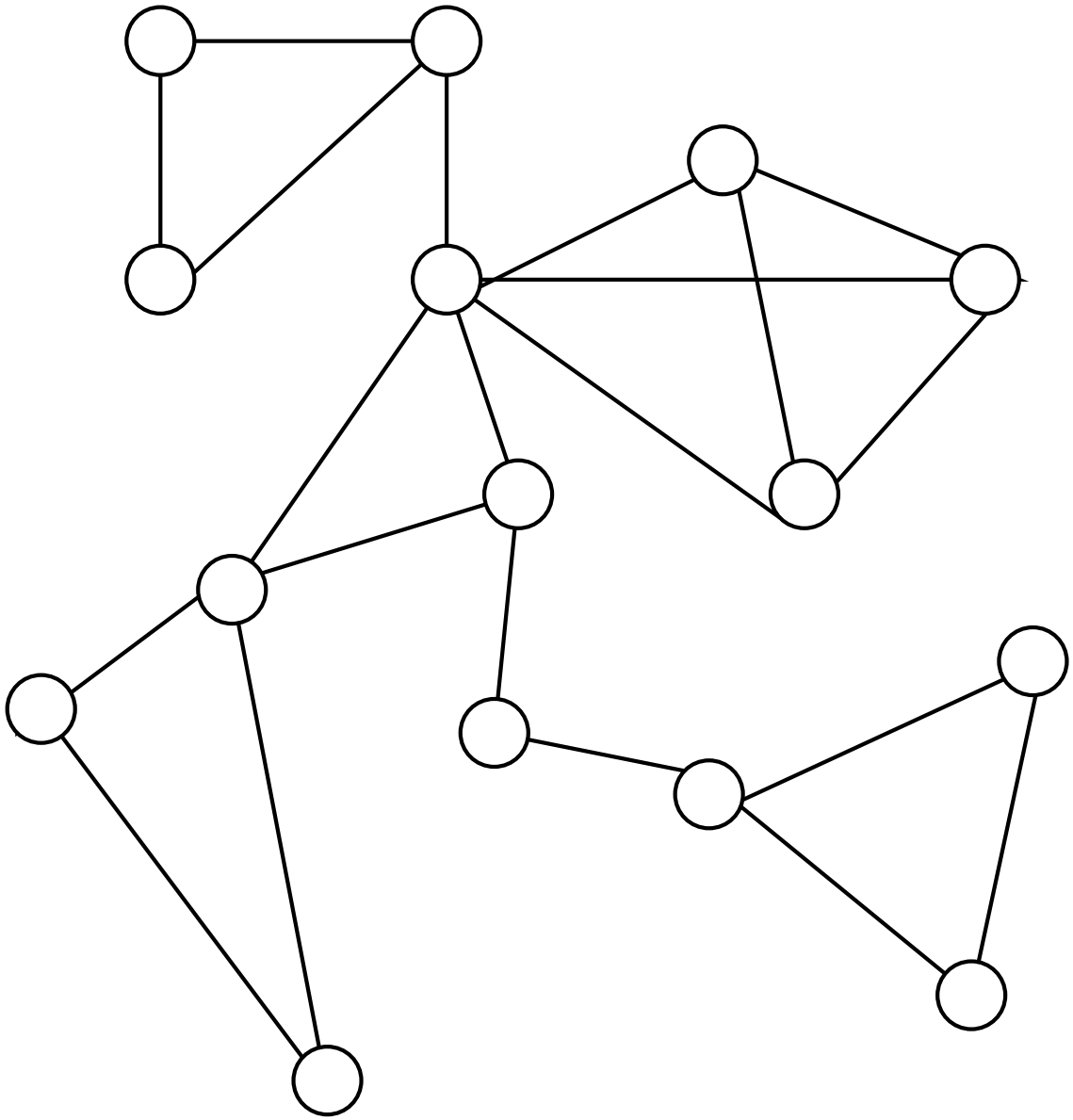


A success

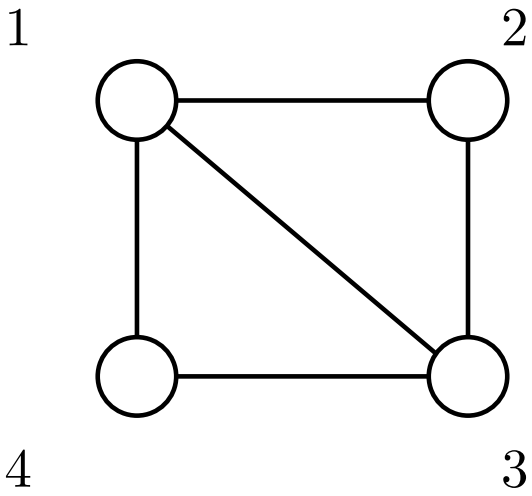


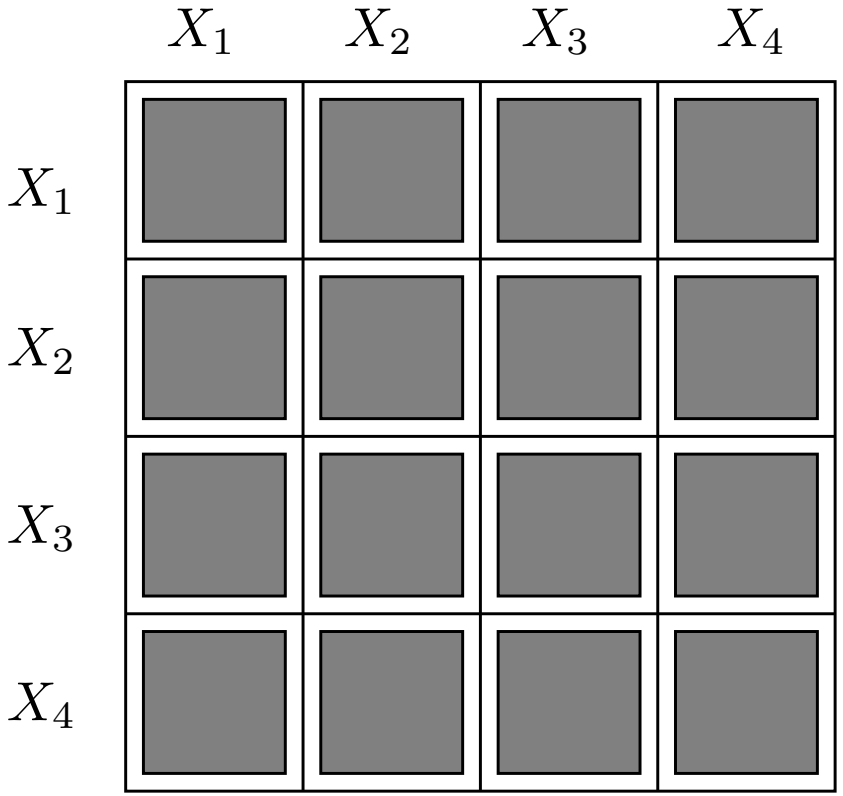


A bigger success: Dinosaur graph

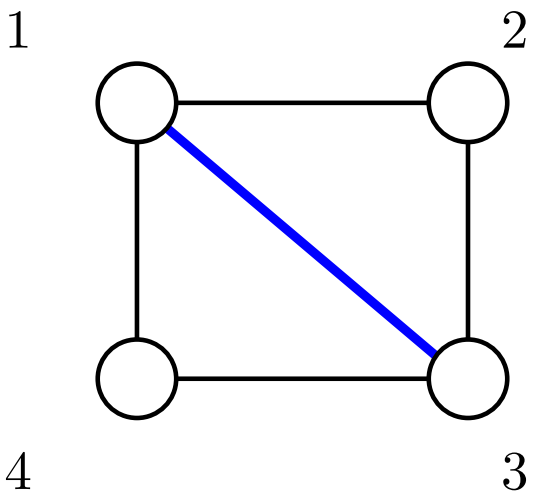


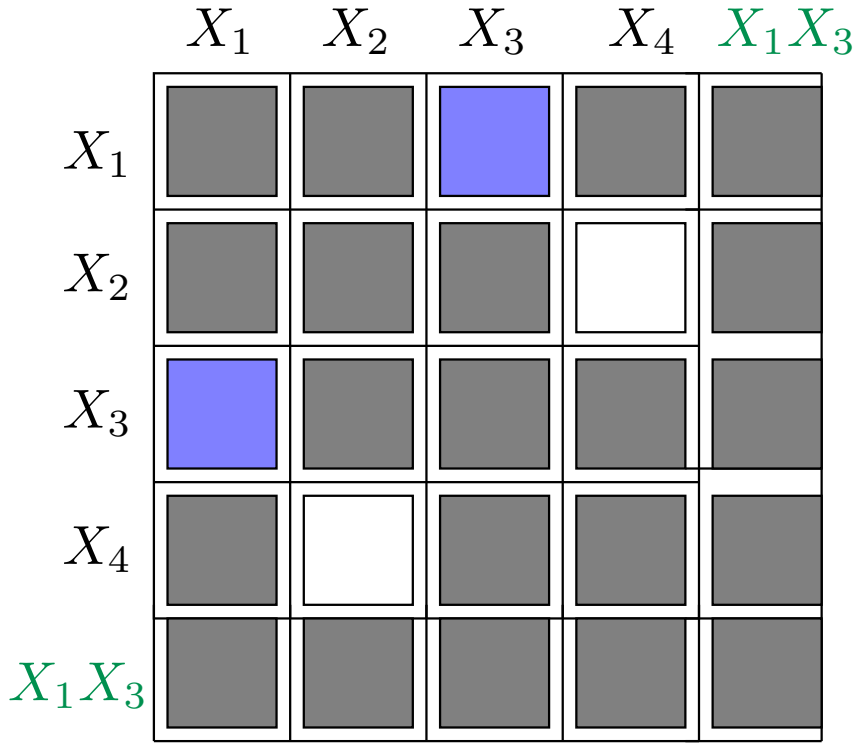
A triangulated counter example





Generalized moment matrices: a success





Generalized moment matrices

- discrete graphical model with $x_j \in \mathcal{X} = \{0, 1, \ldots, m-1\}$:
- for each clique C, define subset of possible configurations

$$\mathcal{X}_0^{|C|} := \big\{ j_C = (j_s, s \in C) \mid j_s \neq 0 \quad \text{for all } s \in C \big\}.$$

- parameterization as a minimal exponential family:

$$\mathbb{Q}_\theta(x_1, \ldots, x_p) = \exp\Big\{ \sum_{C \in \mathcal{C}} \sum_{J \in \mathcal{X}_0^{|C|}} \theta_{C;J} \mathbb{I}\left[x_C = J\right] - A(\theta) \Big\},$$

- for any set $\mathcal{S}$ of subsets of vertices, define random vector:

$$\Phi(X; \mathcal{S}) = \big\{ \mathbb{I}\left[X_C = J\right],\ J \in \mathcal{X}_0^{|C|}, \quad S \in \mathcal{S} \big\}.$$

Generalized inverse covariances and graph structure

- for any set $\mathcal{S}$ of subsets of vertices, define random vector:

$$\Phi(X; \mathcal{S}) = \left\{ \mathbb{I}\left[X_C = J\right],\ J \in \mathcal{X}_0^{|C|}, \quad S \in \mathcal{S} \right\}.$$

- for any set of subsets $\mathcal{S}$, define

$$\mathrm{pow}(\mathcal{S}) = \left\{ \mathrm{pow}(S), |\ \ S \in \mathcal{S} \right\}.$$

Theorem (Loh & Wainwright, 2012)

For any clique set $\mathcal{T}$ in some triangulation of G, the inverse Γ of the augmented covariance matrix $\mathrm{cov}(\Phi(X; pow(\mathcal{T})))$ is block graph-structured in the following sense:

(a) *For any two subsets A and B which are not subsets of the same maximal clique, the block $\Gamma(pow(A), pow(B))$ is zero.*

(b) *For a generic parameter vector θ, the block $\Gamma(pow(A), pow(B))$ is non-zero whenever A and B belong to a common maximal clique.*

Separator sets are sufficient

- consider any triangulated version $\widetilde{G} = (V, \widetilde{E})$
- let $\mathcal{S}$ be set of separator sets in $\widetilde{G}$

Corollary

Let Γ *be the inverse of* $\mathrm{cov}(\Phi(X; V \cup pow(\mathcal{S})))$. *Then* $\Gamma_{V,V}$ *is block edge-structured, meaning that*

$$(s,t) \notin \widetilde{E} \quad \Longleftrightarrow \quad \Gamma_{s,t} = 0.$$

Corollary

For any graph with singleton separator sets, the inverse covariance matrix is graph-structured. (Includes forests/trees as a special case).

Proof ideas

Key ingredients:

- Legendre duality and minimal exponential families (e.g., Brown, 1986)
- Junction tree representation (e.g., Lauritzen & Spielgelhalter, 1988)
- Entropy decompositions and marginal polytopes (Wainwright & Jordan, 2008)

Proof ideas

Key ingredients:

- Legendre duality and minimal exponential families (e.g., Brown, 1986)
- Junction tree representation (e.g., Lauritzen & Spielgelhalter, 1988)
- Entropy decompositions and marginal polytopes (Wainwright & Jordan, 2008)
- Cumulant generating function of our exponential family:

$$A(\theta) := \log \Big\{ \sum_{x} \exp \Big(\sum_{C \in \mathcal{C}} \sum_{j_C \in \mathcal{X}_0^{|C|}} \theta_{C;J_C} \mathbb{I}\left[x_C = J_C\right] \Big) \Big\}.$$

Proof ideas

Key ingredients:

- Legendre duality and minimal exponential families (e.g., Brown, 1986)
- Junction tree representation (e.g., Lauritzen & Spielgelhalter, 1988)
- Entropy decompositions and marginal polytopes (Wainwright & Jordan, 2008)
- Cumulant generating function of our exponential family:

$$A(\theta) := \log \Big\{ \sum_{x} \exp \Big(\sum_{C \in \mathcal{C}} \sum_{j_C \in \mathcal{X}_0^{|C|}} \theta_{C;J_C} \mathbb{I}\left[x_C = J_C\right] \Big) \Big\}.$$

- strictly convex with second partial derivatives

$$\frac{\partial A}{\partial \theta_{C;J_C} \theta_{D;K_D}}(\theta) = \operatorname{cov} \left\{ \mathbb{I}\left[X_C = J_C\right], \, \mathbb{I}\left[X_D = K_D\right] \right\}.$$

- result concerns the structure of $(\nabla^2 A(\theta))^{-1}$: connect this to the Legendre dual A^* of A.

Some statistical consequences

- regularized log-determinant methods for discrete graphical models (Banerjee, d'Aspremont & El Ghaoui, 2009):
  - consistent for "good" graphs (i.e., singleton separator sets, including trees)
  - inconsistent in general; partial consistency results are possible

Some statistical consequences

- regularized log-determinant methods for discrete graphical models (Banerjee, d'Aspremont & El Ghaoui, 2009):
  - consistent for "good" graphs (i.e., singleton separator sets, including trees)
  - inconsistent in general; partial consistency results are possible

- local methods for discrete graphical selection (Loh & W., 2012)
  - for "good" graph, ordinary nodewise linear regression is consistent:

$$\widehat{\theta}[s] \in \arg\min_{\theta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2} \|X_s - X_{\setminus\{s\}}\theta\|_2^2 + \lambda_n^2 \|\theta\|_1^2 \right\}.$$

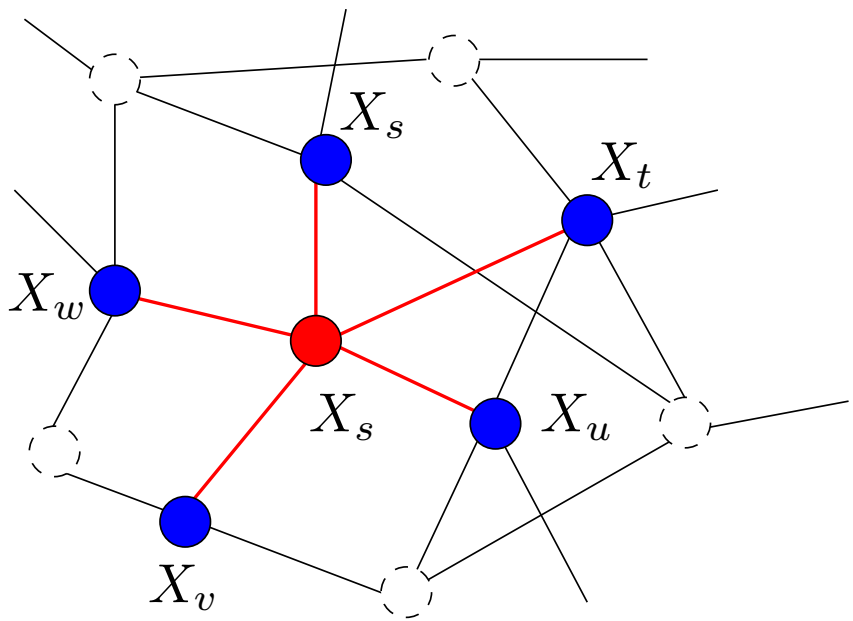
  - for graphs with maximum degree d, nodewise regression with lifted moment matrices of order d are consistent

Markov property and neighborhood structure

- Markov properties encode neighborhood structure:

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

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



- pseudolikelihood method (Besag, 1974)
- basis of many graph learning algorithms (Spirtes et al., 2000; Abeel et al., 2006; Meinshausen & Buhlmann, 2006)

Two methods for discrete binary models

Method 1: ℓ_1-regularized logistic regression

- perform **logistic** regression of each variable X_s on $X_{V \setminus \{s\}}$
- ℓ_1-regularization to enforce sparsity
- known to succeed w.h.p. with sample size $n \gtrsim d^2 \log p$, where d is maximum degree (Ravikumar et al., 2010)

Two methods for discrete binary models

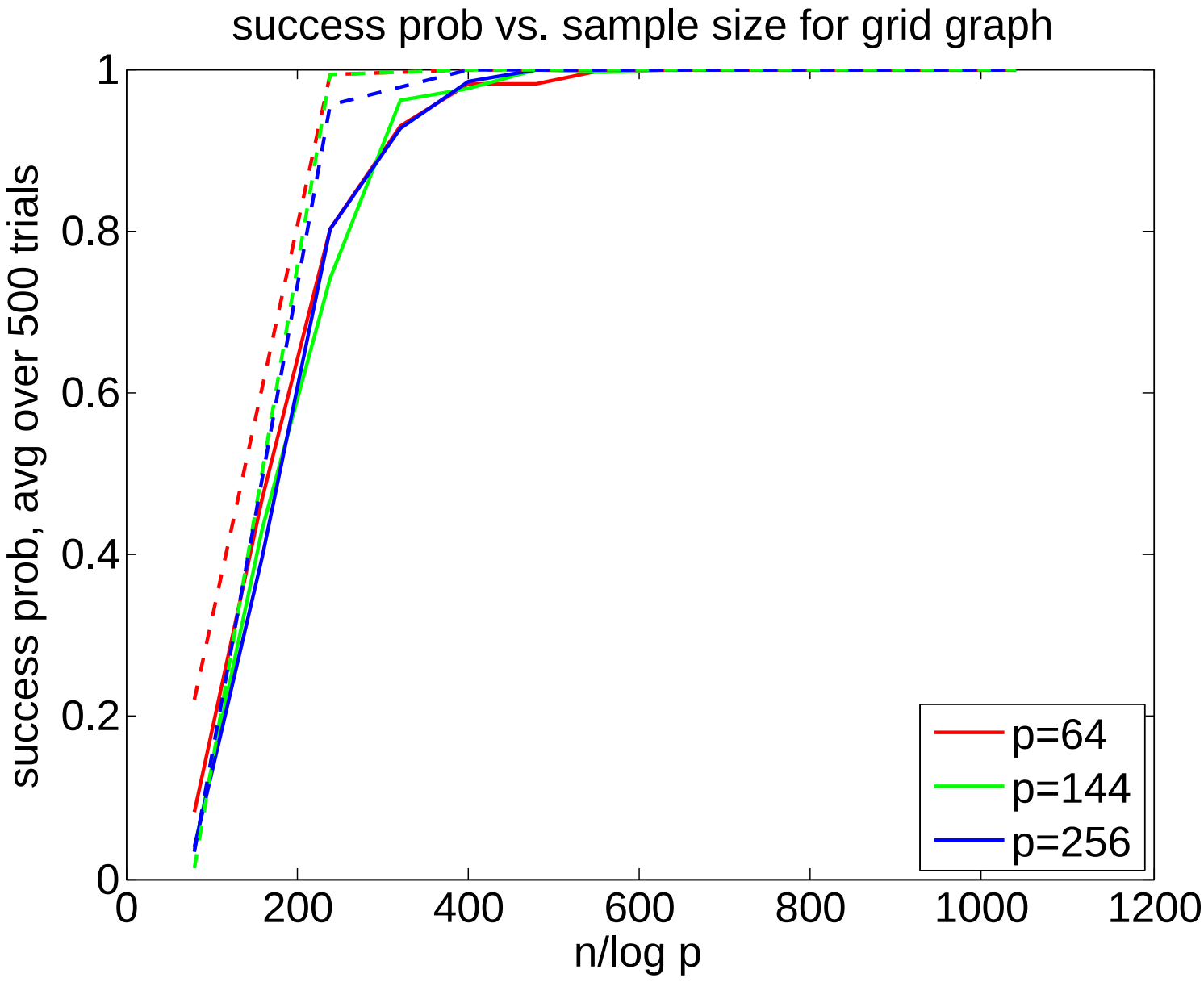
Method 1: ℓ_1-regularized logistic regression

- perform **logistic** regression of each variable X_s on $X_{V \setminus \{s\}}$
- ℓ_1-regularization to enforce sparsity
- known to succeed w.h.p. with sample size $n \gtrsim d^2 \log p$, where d is maximum degree (Ravikumar et al., 2010)

Method 2: ℓ_1-regularized **linear** regression

- perform linear regression of each variable X_s on the d-extended version of $X_{V \setminus \{s\}}$
- ℓ_1-regularization to enforce sparsity
- involves $\mathcal{O}(p^d)$ covariates, so suitable only for bounded degree graphs
- extends naturally to missing/corrupted covariates

Logistic versus linear regression for grid



Straightforward extension to missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$ of matrix $X_{\setminus\{s\}} \in \mathbb{R}^{n \times (p-1)}$:

$$\widetilde{Z}_{ij} = \begin{cases} X_{ij} & \text{with probability } 1 - \alpha \\ \star & \text{with probability } \alpha. \end{cases}$$

Straightforward extension to missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$ of matrix $X_{\setminus\{s\}} \in \mathbb{R}^{n \times (p-1)}$:

$$\widetilde{Z}_{ij} = \begin{cases} X_{ij} & \text{with probability } 1 - \alpha \\ \star & \text{with probability } \alpha. \end{cases}$$

- set $\star \equiv 0$ and $\widehat{Z} = \frac{\widetilde{Z}}{1-\alpha}$, and form quantities

$$\widehat{\Gamma} \quad = \quad \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \, \mathrm{diag}\left(\frac{\widehat{Z}^T \widehat{Z}}{n}\right), \quad \text{and} \quad \widehat{\gamma} \quad = \quad \frac{\widehat{Z}^T y}{n},$$

- unbiased estimates of $\Gamma = \mathrm{cov}(X_{\setminus\{s\}})$ and $\gamma = \mathrm{cov}(X_s, X_{\setminus\{s\}})$

Straightforward extension to missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$ of matrix $X_{\setminus\{s\}} \in \mathbb{R}^{n \times (p-1)}$:

$$\widetilde{Z}_{ij} = \begin{cases} X_{ij} & \text{with probability } 1 - \alpha \\ \star & \text{with probability } \alpha. \end{cases}$$

- set $\star \equiv 0$ and $\widehat{Z} = \frac{\widetilde{Z}}{1-\alpha}$, and form quantities

$$\widehat{\Gamma} \quad = \quad \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \, \mathrm{diag}\left(\frac{\widehat{Z}^T \widehat{Z}}{n}\right), \quad \text{and} \quad \widehat{\gamma} \quad = \quad \frac{\widehat{Z}^T y}{n},$$

- unbiased estimates of $\Gamma = \mathrm{cov}(X_{\setminus\{s\}})$ and $\gamma = \mathrm{cov}(X_s, X_{\setminus\{s\}})$

- solve optimization problem:

$$\widehat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta - \langle \widehat{\gamma}, \, \theta \rangle + \lambda_n^2 \|\theta\|_1^2 \right\}$$

Straightforward extension to missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$ of matrix $X_{\setminus\{s\}} \in \mathbb{R}^{n \times (p-1)}$:

$$\widetilde{Z}_{ij} = \begin{cases} X_{ij} & \text{with probability } 1 - \alpha \\ \star & \text{with probability } \alpha. \end{cases}$$

- set $\star \equiv 0$ and $\widehat{Z} = \frac{\widetilde{Z}}{1-\alpha}$, and form quantities

$$\widehat{\Gamma} \quad = \quad \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \, \mathrm{diag}\left(\frac{\widehat{Z}^T \widehat{Z}}{n}\right), \quad \text{and} \quad \widehat{\gamma} \quad = \quad \frac{\widehat{Z}^T y}{n},$$

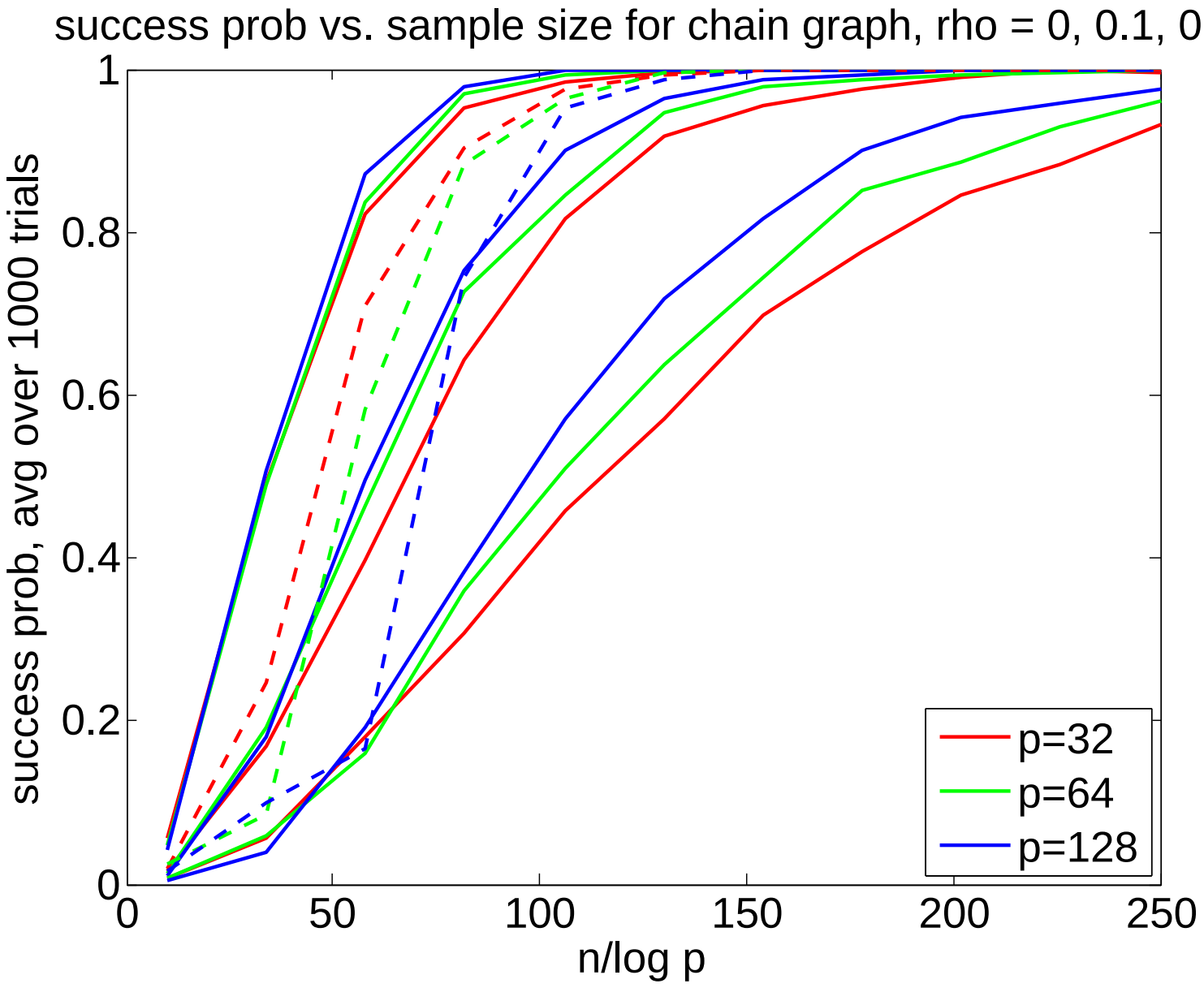
- unbiased estimates of $\Gamma = \mathrm{cov}(X_{\setminus\{s\}})$ and $\gamma = \mathrm{cov}(X_s, X_{\setminus\{s\}})$

- solve optimization problem:

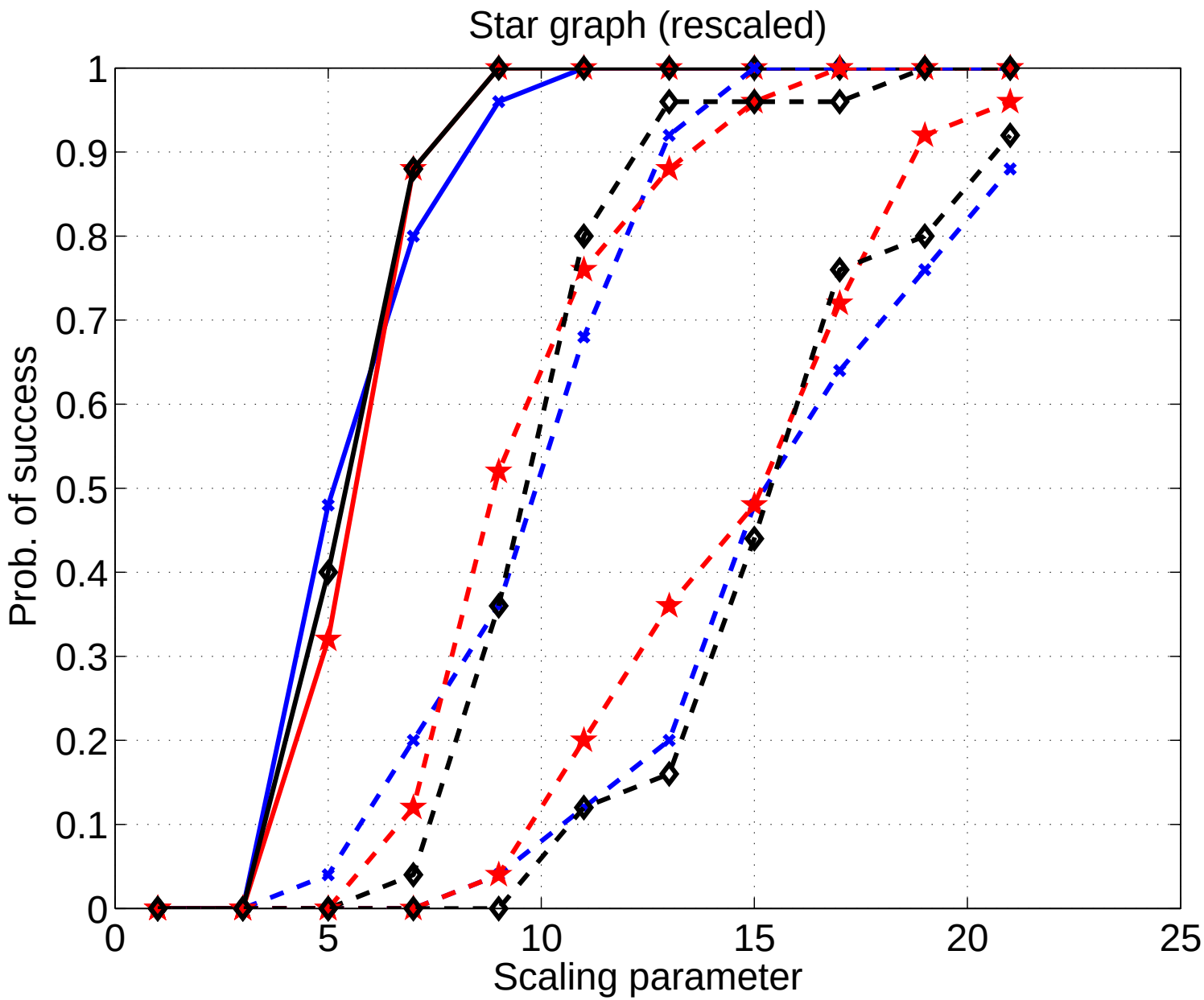
$$\widehat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta - \langle \widehat{\gamma}, \, \theta \rangle + \lambda_n^2 \|\theta\|_1^2 \right\}$$

- "corrected" estimates of this type:
  - Xu & You (2007): corrected Lasso in classical setting
  - Loh & Wainwright (2012): corrected Lasso and algorithms for non-convexity

Performance with missing data: Chain graph



Performance with missing data: Star graph



Statistical guarantees

- take n independent samples $\{X_i\}_{i=1}^n$ from Ising model with zero mean, and inverse covariance matrix $\Theta^* = (\mathrm{cov}(X))^{-1}$
- entries of samples $\{X_i\}_{i=1}^n$ missing independently with probability α
- graph has maximum degree d, and singleton separator sets

Statistical guarantees

- take n independent samples $\{X_i\}_{i=1}^n$ from Ising model with zero mean, and inverse covariance matrix $\Theta^* = (\mathrm{cov}(X))^{-1}$
- entries of samples $\{X_i\}_{i=1}^n$ missing independently with probability α
- graph has maximum degree d, and singleton separator sets

Theorem (Loh & W., 2012)

As long as $n \gtrsim \frac{d^2 \log p}{(1-\alpha)^2}$ *and* $\lambda_n \asymp \sqrt{\frac{\log p}{n}}$, *then with probability greater than* $1 - 2\exp(-cn\lambda_n^2)$, *any global optimum* $\widehat{\theta}$ *satisfies the bound*

$$\|\widehat{\theta} - \theta^*\|_\infty \lesssim |||\Theta^*|||_\infty \sqrt{\frac{\log p}{n}}.$$

Consequently, the method is graph-selection consistent as long as $\min_{t \in \mathcal{N}(s)} |\theta^*_{st}| \gtrsim \sqrt{\frac{\log p}{n}}$.

Summary

- graphical model selection:
  - wide-range of applications
  - convex relaxations play an important role
  - Gaussian data benefits from link to inverse covariance matrix

- general connection between graph structure and inverse covariance
  - depends on structure of separator sets in a junction tree
  - inverses of higher-order moment matrices are useful

- various statistical consequences for discrete graph selection

Summary

- graphical model selection:
  - wide-range of applications
  - convex relaxations play an important role
  - Gaussian data benefits from link to inverse covariance matrix

- general connection between graph structure and inverse covariance
  - depends on structure of separator sets in a junction tree
  - inverses of higher-order moment matrices are useful

- various statistical consequences for discrete graph selection

- many remaining questions:
  - how to handle more realistic models of missing data?
  - provable guarantees for other types of non-convex M-estimators?
  - discrete graphical models with hidden variables?

Some papers/pre-prints

- Agarwal, Negahban & W. (2012). Fast convergence rates for high-dimensional statistical recovery. *Annals of Statistics*, 40(5): 2452–2482, December.

- Loh & W. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3): 1637–1664, September.

- Loh & W. (2012). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses *Arxiv pre-print*.