Learning in graphical models: Missing data and rigorous guarantees with non-convexity

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Based on joint work with:

John Lafferty (CMU)
Po-Ling Loh (UC Berkeley)
Pradeep Ravikumar (UT Austin)

Introduction

- Markov random fields (undirected graphical models): central in many application areas of science/engineering:

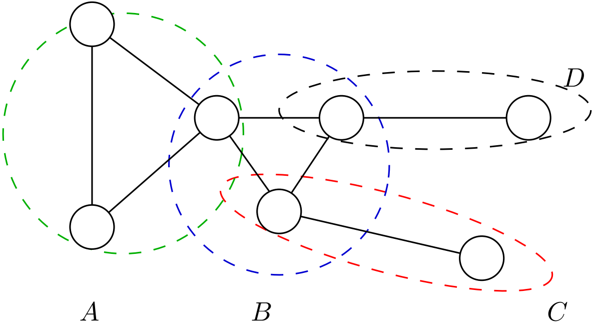
Introduction

- Markov random fields (undirected graphical models): central in many application areas of science/engineering:

- some fundamental problems
  - *counting/integrating:* computing marginal distributions and partition functions
  - *optimization:* computing most probable configurations (or top M-configurations)
  - *graph learning:* fitting and selecting models on the basis of data

Graph structure and factorization

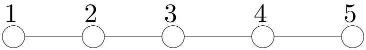
- Markov random field: random vector $(X_1, \ldots, X_p)$ with distribution factoring according to a graph $G = (V, E)$:

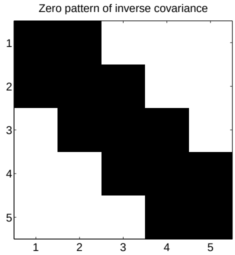


- Hammersley-Clifford theorem: factorization over cliques

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

Some pairwise graphical models





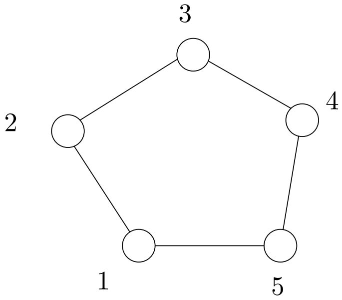
- $p \times p$ matrix of weights $\Theta = [\theta_{st}]$
- Ising model $(X_1, \ldots, X_p) \in \{0, 1\}^p$:

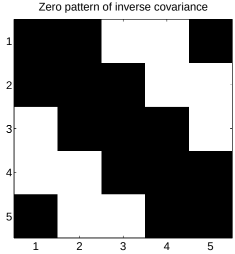
$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) = \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}.$$

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

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

Some pairwise graphical models





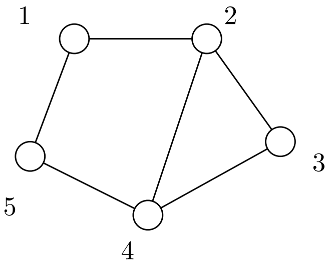
- $p \times p$ matrix of weights $\Theta = [\theta_{st}]$
- Ising model $(X_1, \ldots, X_p) \in \{0, 1\}^p$:

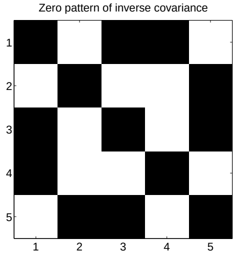
$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) = \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}.$$

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

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

Some pairwise graphical models





- $p \times p$ matrix of weights $\Theta = [\theta_{st}]$
- Ising model $(X_1, \ldots, X_p) \in \{0, 1\}^p$:

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) = \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}.$$

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

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

Graphical model learning

- drawn n samples from

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) = \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s^2 + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- graph G and matrix $[\Theta]_{st} = \theta_{st}$ of edge weights are unknown

Graphical model learning

- drawn n samples from

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s^2 + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- graph G and matrix $[\Theta]_{st} = \theta_{st}$ of edge weights are unknown

- data matrix $\mathbf{X}_1^n \in \{0,1\}^{n \times p}$ (or in $\mathbf{X}_1^n \in \mathbb{R}^{n \times p}$)
- estimator $\mathbf{X}_1^n \mapsto \widehat{\Theta}$

Graphical model learning

- drawn n samples from

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s^2 + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- graph G and matrix $[\Theta]_{st} = \theta_{st}$ of edge weights are unknown

- data matrix $\mathbf{X}_1^n \in \{0,1\}^{n \times p}$ (or in $\mathbf{X}_1^n \in \mathbb{R}^{n \times p}$)
- estimator $\mathbf{X}_1^n \mapsto \widehat{\Theta}$

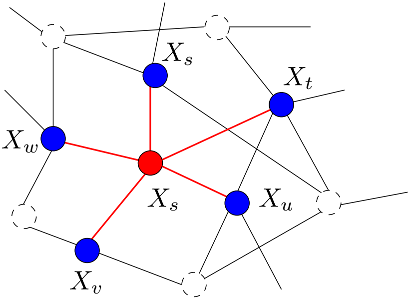
- various loss functions are possible:
  - graph selection: $\text{supp}[\widehat{\Theta}] = \text{supp}[\Theta]$?
  - bounds on Kullback-Leibler divergence $D(\mathbb{Q}_{\widehat{\Theta}} \parallel \mathbb{Q}_{\Theta})$
  - bounds on $\|\widehat{\Theta} - \Theta\|_{\text{op}}$.

Markov property and neighborhood structure

- Markov properties encode neighborhood structure:

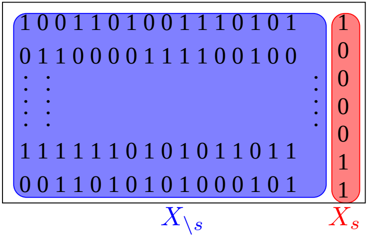
$$\underbrace{({\color{red}X_s} \quad | \quad X_{V\backslash s})}_{\text{Condition on full graph}} \stackrel{d}{=} \underbrace{({\color{red}X_s} \quad | \quad {\color{blue}X_{N(s)}})}_{\text{Condition on Markov blanket}}$$

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



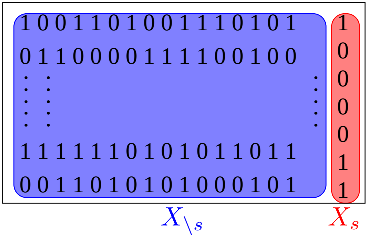
- basis of pseudolikelihood method (Besag, 1974)
- basis of many graph learning algorithm (Friedman et al., 1999; Csiszar & Talata, 2005; Abeel et al., 2006; Meinshausen & Buhlmann, 2006)

Graph selection via neighborhood regression



Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

Graph selection via neighborhood regression

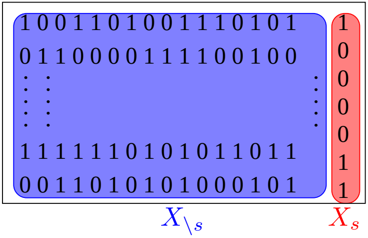


Predict X_s based on $X_{\backslash s} := \{X_s,\ t \neq s\}$.

1. For each node $s \in V$, compute (regularized) max. likelihood estimate:

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

Graph selection via neighborhood regression



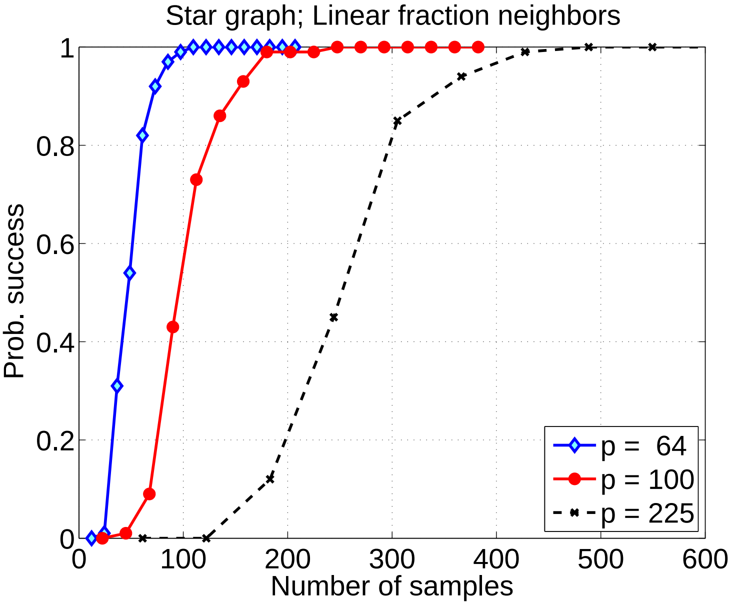
Predict X_s based on $X_{\backslash s} := \{X_s,\ t \neq s\}$.

1. For each node $s \in V$, compute (regularized) max. likelihood estimate:

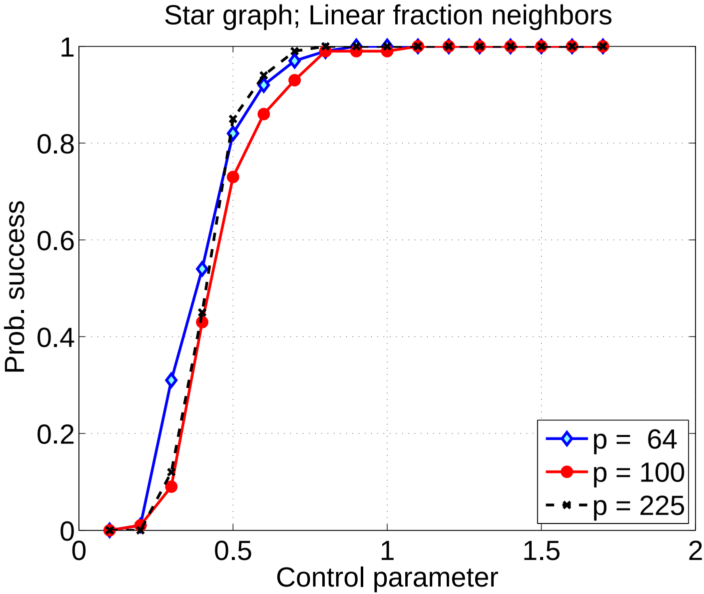
$$\widehat{\theta}[s] \quad := \quad \arg\min_{\theta \in \mathbb{R}^{p-1}} \Big\{ \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; X_{\backslash s}^{(i)})}_{\text{local log. likelihood}} \quad + \quad \lambda_n \underbrace{\|\theta\|_1}_{\text{regularization}} \Big\}$$

2. Estimate the local neighborhood $\widehat{N}(s)$ as support of regression vector $\widehat{\theta}[s] \in \mathbb{R}^{p-1}$.

Empirical behavior: Unrescaled plots



Empirical behavior: Appropriately rescaled



Sufficient conditions for consistent Ising 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 (Ravikumar, W. & Lafferty, 2006, 2010)

Sufficient conditions for consistent Ising 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 (Ravikumar, W. & Lafferty, 2006, 2010)

Under incoherence conditions, for a rescaled sample

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

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

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

Sufficient conditions for consistent Ising 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 (Ravikumar, W. & Lafferty, 2006, 2010)

Under incoherence conditions, for a rescaled sample

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

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

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

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

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)

- ℓ_1-based method: sharper achievable rates, also failure for θ large enough to violate incoherence (Bento & Montanari, 2009)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)

- ℓ_1-based method: sharper achievable rates, also failure for θ large enough to violate incoherence (Bento & Montanari, 2009)

- empirical study: ℓ_1-based method can succeed beyond phase transition on Ising model (Aurell & Ekeberg, 2011)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)
- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)
- ℓ_1-based method: sharper achievable rates, also failure for θ large enough to violate incoherence (Bento & Montanari, 2009)
- empirical study: ℓ_1-based method can succeed beyond phase transition on Ising model (Aurell & Ekeberg, 2011)
- simpler neighborhood-based methods: thresholding, mutual information, greedy-methods
  - Anandkumar, Tan & Willsky, 2010a, 2010b
  - Netrapalli et al., 2010
- refined dependence on graph structure (Anandkumar et al; talk later today)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)

- ℓ_1-based method: sharper achievable rates, also failure for θ large enough to violate incoherence (Bento & Montanari, 2009)

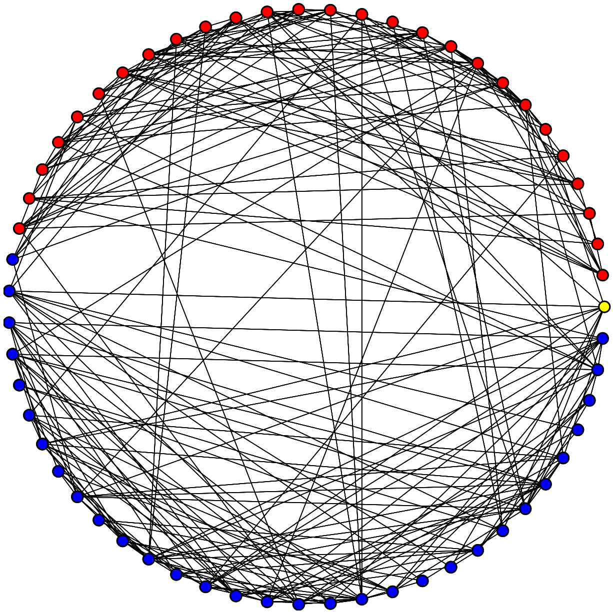
- empirical study: ℓ_1-based method can succeed beyond phase transition on Ising model (Aurell & Ekeberg, 2011)

- simpler neighborhood-based methods: thresholding, mutual information, greedy-methods
  - Anandkumar, Tan & Willsky, 2010a, 2010b
  - Netrapalli et al., 2010

- refined dependence on graph structure (Anandkumar et al; talk later today)

- "list-decoding" for graphical models (Vats & Moura, 2011)

US Senate network (2004–2006 voting)



A challenge

The reality:

In practice, samples $X = (X_1, \ldots, X_p)$ are not perfectly observed.

A challenge

The reality:

In practice, samples $X = (X_1, \ldots, X_p)$ are not perfectly observed.

- Examples:
  - Missing data (e.g., voting records):

$$\begin{bmatrix} X_1 & X_2 & X_3 & X_4 & \ldots & X_p \end{bmatrix} = \begin{bmatrix} 0 & 1 & * & 1 & \ldots & 0 \end{bmatrix}.$$

  - Noisy and corrupted data:

$$Z = X + W$$

A challenge

The reality:

In practice, samples $X = (X_1, \ldots, X_p)$ are not perfectly observed.

- Examples:
  - Missing data (e.g., voting records):

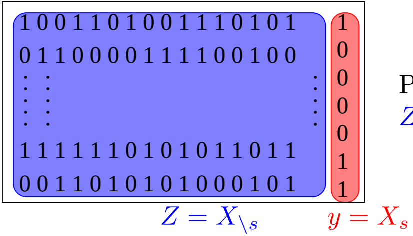
$$\begin{bmatrix} X_1 & X_2 & X_3 & X_4 & \ldots & X_p \end{bmatrix} = \begin{bmatrix} 0 & 1 & * & 1 & \ldots & 0 \end{bmatrix}.$$

  - Noisy and corrupted data:

$$Z = X + W$$

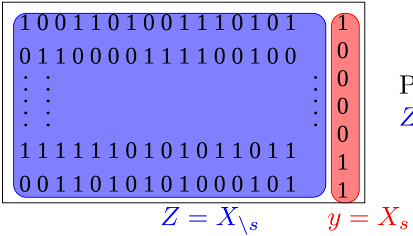
- standard methods for missing data (e.g., EM algorithm) lead to non-convex problems
- very difficult to provide rigorous guarantees

Gaussian case (linear regression)



Predict $y = X_s$ based on other variables $Z = X_{\setminus s} := \{X_s,\ t \neq s\}$.

Gaussian case (linear regression)

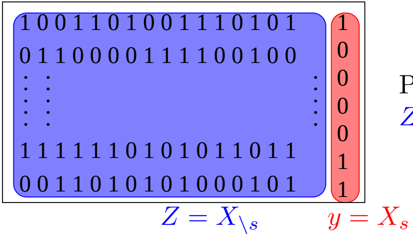


Predict $y = X_s$ based on other variables $Z = X_{\setminus s} := \{X_s,\ t \neq s\}$.

- when (y, Z) is fully observed, solve problem

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \frac{1}{2n} \|y - Z\theta\|_2^2 + \lambda_n \|\theta\|_1 \Big\}$$

Gaussian case (linear regression)

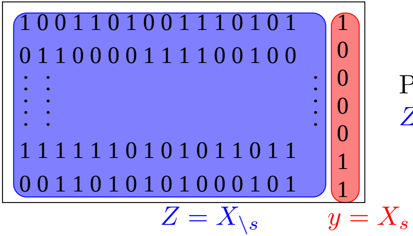


Predict $y = X_s$ based on other variables $Z = X_{\backslash s} := \{X_s,\ t \neq s\}$.

- when (y, Z) is fully observed, solve problem

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta - \langle \widehat{\gamma},\, \theta \rangle + \lambda \|\theta\|_1 \Big\} \quad \text{where } \widehat{\Gamma} = \tfrac{Z^T Z}{n} \text{ and } \widehat{\gamma} = \tfrac{Z^T y}{n}.$$

Gaussian case (linear regression)



Predict $y = X_s$ based on other variables $Z = X_{\setminus s} := \{X_s,\ t \neq s\}$.

- when (y, Z) is fully observed, solve problem

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta - \langle \widehat{\gamma},\, \theta \rangle + \lambda \|\theta\|_1 \Big\} \quad \text{where } \widehat{\Gamma} = \frac{Z^T Z}{n} \text{ and } \widehat{\gamma} = \frac{Z^T y}{n}.$$

- more general family of estimators: let $(\widehat{\Gamma}, \widehat{\gamma})$ be any unbiased estimators of

$$\operatorname{cov}(Z_i) \in \mathbb{R}^{(p-1)\times(p-1)} \quad \text{and} \quad \operatorname{cov}(y_i Z_i) \in \mathbb{R}^{p-1}.$$

Example: Estimator for missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$

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

Example: Estimator for missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$

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

- Natural unbiased estimates: set $\star \equiv 0$ and $\widehat{Z} := \frac{\widetilde{Z}}{(1-\alpha)}$:

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

Example: Estimator for missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$

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

- Natural unbiased estimates: set $\star \equiv 0$ and $\widehat{Z} := \frac{\widetilde{Z}}{(1-\alpha)}$:

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

- solve optimization problem: $\widehat{\theta} \in \arg\min_\theta \big\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta - \langle \widehat{\gamma}, \theta \rangle + \lambda \|\theta\|_1 \big\}$.

Example: Estimator for missing data

- observe corrupted version $\widetilde{Z} \in \mathbb{R}^{n \times (p-1)}$

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

- Natural unbiased estimates: set $\star \equiv 0$ and $\widehat{Z} := \frac{\widetilde{Z}}{(1-\alpha)}$:

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

- solve optimization problem: $\widehat{\theta} \in \arg\min_\theta \big\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta - \langle \widehat{\gamma}, \, \theta \rangle + \lambda \|\theta\|_1 \big\}$.

Challenge:

Matrix $\widehat{\Gamma}$ not positive semidefinite $\Longrightarrow$ non-convex program.

Theoretical guarantees on statistical error

- take n i.i.d. samples multivariate Gaussian in p-dimensions
- missing probability $\alpha \in [0, 1)$
- inverse covariance matrix $\Theta^* \in \mathbb{R}^{p \times p}$:
  - bounded eigenspectrum
  - at most d non-zero entries per row

Theoretical guarantees on statistical error

- take n i.i.d. samples multivariate Gaussian in p-dimensions
- missing probability $\alpha \in [0, 1)$
- inverse covariance matrix $\Theta^* \in \mathbb{R}^{p \times p}$:
  - bounded eigenspectrum
  - at most d non-zero entries per row

Theorem (Loh & W., 2011)

Solve non-convex program with regularization $\lambda_n \gtrsim \sqrt{\frac{\log p}{n}}$*. Then with probability greater than* $1 - c_1 \exp(-n\lambda_n^2)$*:*

(a) *For all* $j \in V$*, any global optimum satisfies* $\|\theta_j - \theta^*\|_2 \precsim \frac{1}{1-\alpha} \sqrt{\frac{d \log p}{n}}$.

Theoretical guarantees on statistical error

- take n i.i.d. samples multivariate Gaussian in p-dimensions
- missing probability $\alpha \in [0, 1)$
- inverse covariance matrix $\Theta^* \in \mathbb{R}^{p \times p}$:
  - bounded eigenspectrum
  - at most d non-zero entries per row

Theorem (Loh & W., 2011)

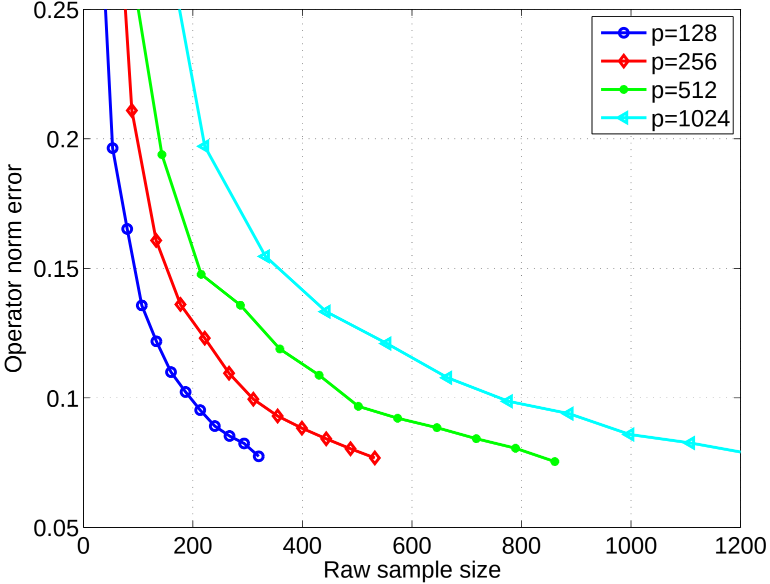
Solve non-convex program with regularization $\lambda_n \gtrsim \sqrt{\frac{\log p}{n}}$. *Then with probability greater than* $1 - c_1 \exp(-n\lambda_n^2)$*:*

(a) *For all* $j \in V$*, any global optimum satisfies* $\|\theta_j - \theta^*\|_2 \precsim \frac{1}{1-\alpha}\sqrt{\frac{d \log p}{n}}$.

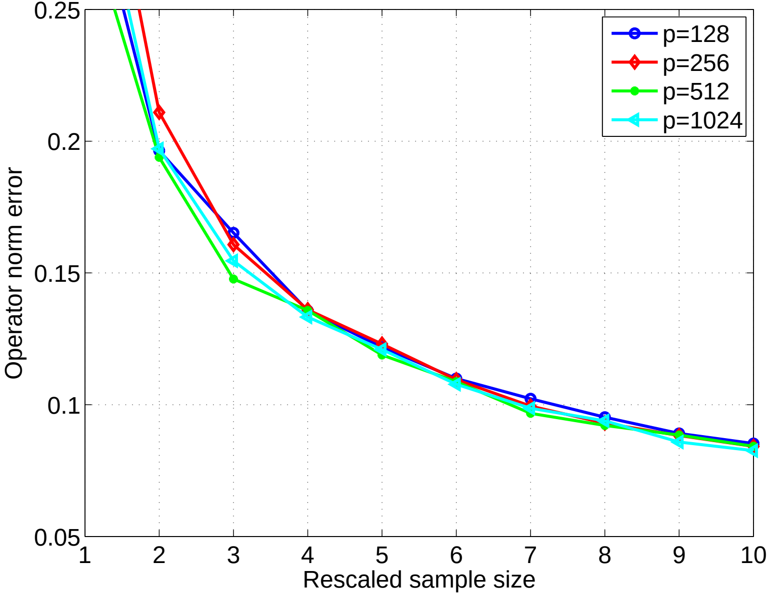
(b) *Combining neighborhood estimates yields a global estimate s.t.:*

$$\|\widehat{\Theta} - \Theta^*\|_{op} \precsim \frac{1}{1-\alpha} \, d\sqrt{\frac{\log p}{n}}.$$

Empirical results (unrescaled)



Empirical results (rescaled)



Projected gradient descent

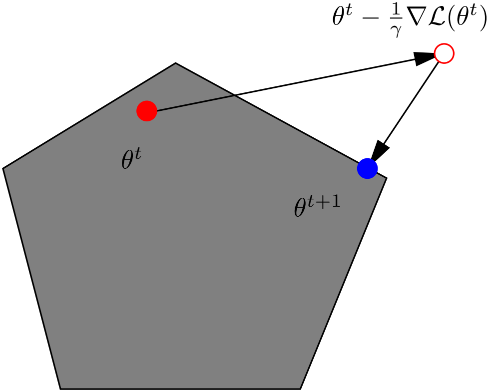
Constrained objective:

$$\widehat{\theta} \in \arg\min_{\theta} \underbrace{\left\{\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i)\right\}}_{\mathcal{L}(\theta)}$$

subject to $\|\theta\|_1 \leq \rho_C$.

With (inverse) stepsize γ:

$$\theta^{t+1} = \Pi_{\rho_c}\big(\theta^t - \frac{1}{\gamma}\nabla\mathcal{L}(\theta^t)\big)$$



- stepsize $\gamma > 0$ related to smoothness of objective function

Convergence for non-convex objective

Theoretical guarantee for non-convex objective

- data drawn from Gaussian graphical model such that:
  - maximum degree d
  - inverse covariance Θ has bounded eigenspectrum
- projected gradient descent with fixed step size: used to estimate row $\theta^* = \Theta_j^* \in \mathbb{R}^p$

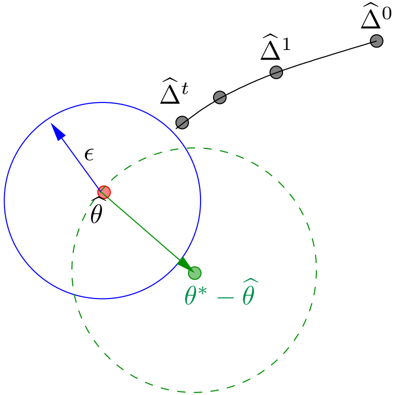
Theorem (Loh & W., 2011)

For $n \gtrsim \frac{d \log p}{(1-\alpha)^2}$, there is w.h.p. a contraction coefficient $\kappa \in (0,1)$ such that for any global optimum $\widehat{\theta}$, the gradient descent iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy

$$\|\theta^t - \widehat{\theta}\|_2^2 \le \kappa^t \underbrace{\|\theta^0 - \widehat{\theta}\|_2^2}_{\text{Opt. error}} + \underbrace{\frac{\log p}{n}\|\widehat{\theta} - \theta^*\|_1^2 + \|\widehat{\theta} - \theta^*\|_2^2}_{\text{Statistical error}}$$

for all iterations $t = 0, 1, 2, \ldots$.

Geometry of result



Optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ decreases geometrically up to statistical tolerance:

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|^2 + o(\underbrace{\|\theta^* - \widehat{\theta}\|^2}_{\text{Statistical error}}) \quad \text{for all iterations } t = 0, 1, 2, \ldots$$

Summary

- graphical model learning: an interesting "inverse" problem
- neighborhood-based approaches:
  - polynomial-time, truly practical
  - match information-theoretic limits up to constant factors

Summary

- graphical model learning: an interesting "inverse" problem
- neighborhood-based approaches:
  - polynomial-time, truly practical
  - match information-theoretic limits up to constant factors

- challenges for {missing, noisy, hidden } data:
  - Gaussian case: non-convex methods have similar guarantees
  - extensions to general variables?
  - combination with fully hidden variables?

Summary

- graphical model learning: an interesting "inverse" problem
- neighborhood-based approaches:
  - polynomial-time, truly practical
  - match information-theoretic limits up to constant factors
- challenges for {missing, noisy, hidden } data:
  - Gaussian case: non-convex methods have similar guarantees
  - extensions to general variables?
  - combination with fully hidden variables?
- geometry of statistical optimization: other guarantees in non-convex settings?

Some papers on graph selection

- Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*.

- Santhanam, P. and Wainwright, M. J. (2008). Information-theoretic limitations of high-dimensional graphical model selection. Presented at *International Symposium on Information Theory,* 2008.

- Loh, P. and Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Arxiv, September 2011.*