# Rank/sparsity minimization and latent variable graphical model selection

**Pablo A. Parrilo**

Laboratory for Information and Decision Systems

Electrical Engineering and Computer Science

Massachusetts Institute of Technology

Joint work with **Venkat Chandrasekaran** and **Alan Willsky**

**Numerical Methods for Continuous Optimization**

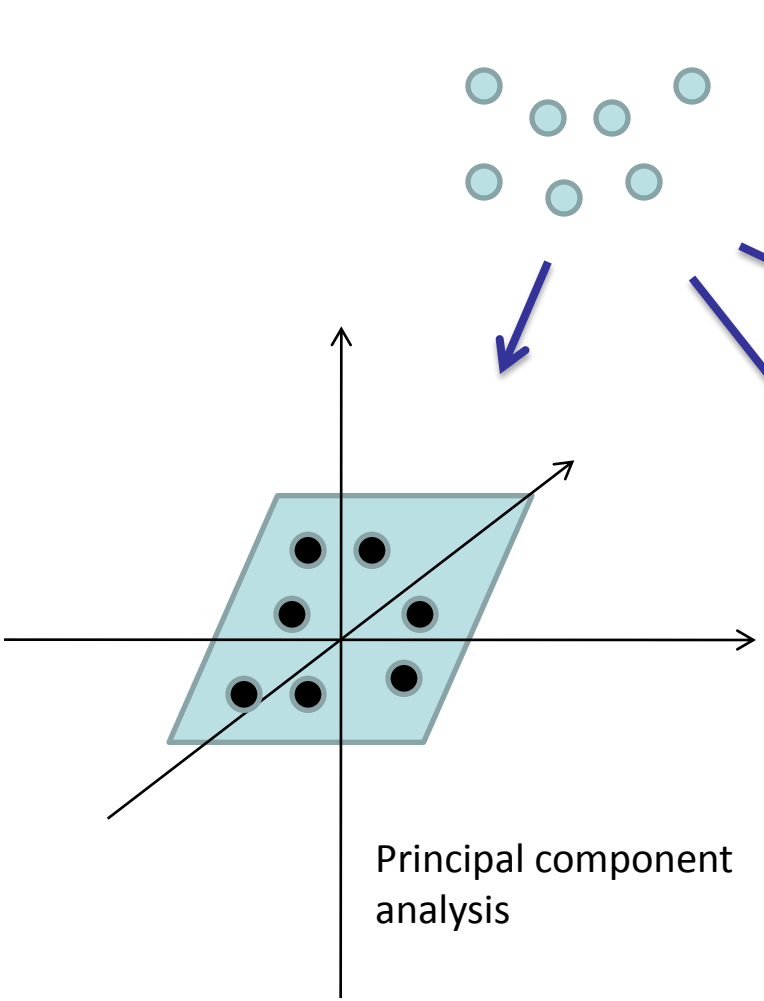
**IPAM – October 2010**

# Overview

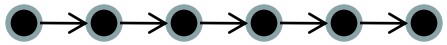
**This talk:**

- Statistical graphical model selection
- Complication: latent variables
- Rank/sparsity decomposition, and generalizations
- Convex optimization formulation
- Identifiability, underlying *geometry*
- Convergence, sample complexity
- Examples and algorithms

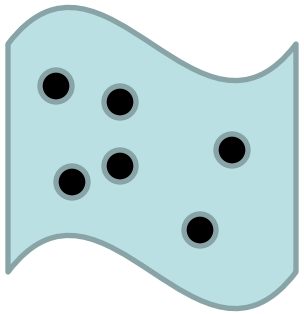
# Statistical model selection



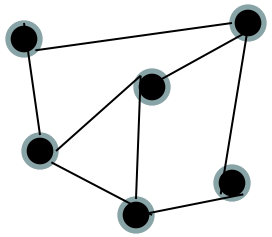
Principal component analysis



Time series



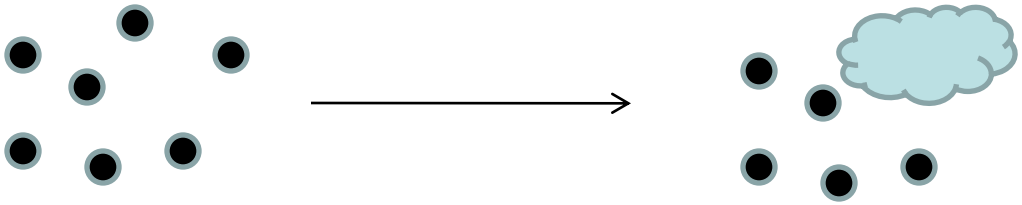
Manifold learning



Graphical models

# Our Problem

- What if some variables are ***not observed?***
  - Don't know *how many* latent variables
  - Don't know *relationship* between observed and latent variables



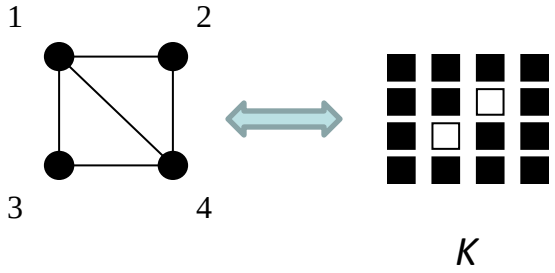
# Gaussian graphical models

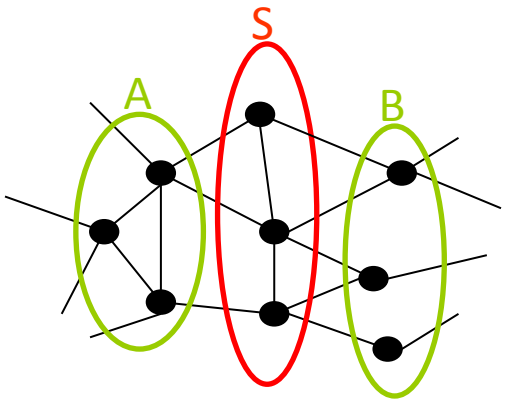
- $p(x) \propto \exp\{-\frac{1}{2}x^T\Sigma^{-1}x\}$
- $\Sigma^{-1} = K$

Covariance matrix

Concentration matrix

- $p(x) \propto \exp\{-\frac{1}{2}x^T K x\}$



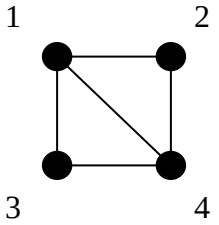

$x_A$ independent of $x_B$ conditioned on $x_S$

- Graphical model: Variables *x* are *Markov* on graph given by *K*

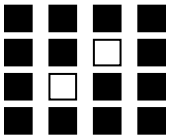
# Latent variable graphical models

$$X \sim \mathcal{N}(0, \Sigma)$$

$X_i$ indep. of $X_j$ cond. on other vars.



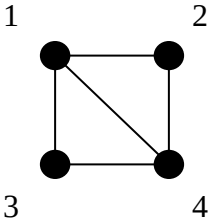

Concentration matrix $\longrightarrow$ $(\Sigma^{-1})_{ij} = 0$

$\Sigma^{-1} =$ 

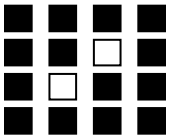
# Latent variable graphical models

$$X \sim \mathcal{N}(0, \Sigma)$$

$X_i$ indep. of $X_j$ cond. on other vars.




Concentration matrix $\longrightarrow$ $(\Sigma^{-1})_{ij} = 0$

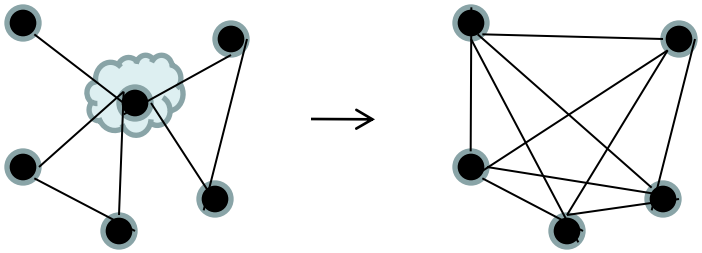
$\Sigma^{-1} =$ 

$$\Sigma = \begin{bmatrix} \Sigma_O & \Sigma_{O,H} \\ \Sigma_{H,O} & \Sigma_H \end{bmatrix} \qquad \Sigma^{-1} = K = \begin{bmatrix} K_O & K_{O,H} \\ K_{H,O} & K_H \end{bmatrix}$$

$$(\Sigma_O)^{-1} = K_O - K_{O,H} K_H^{-1} K_{H,O}$$

Sparse

Low-rank

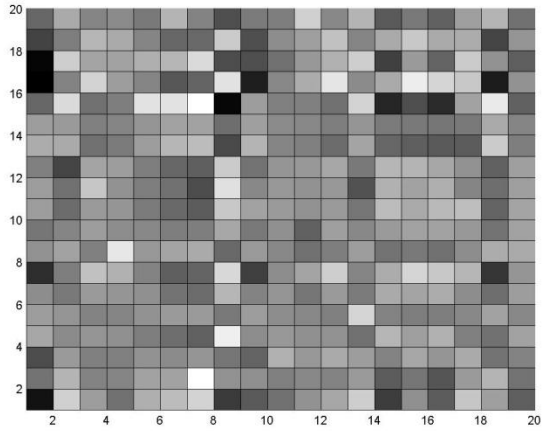


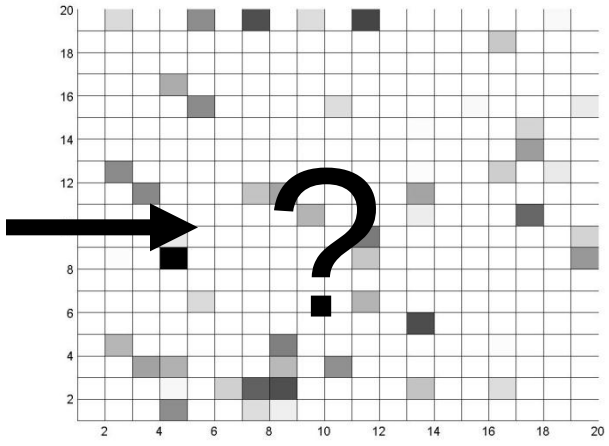
# Proposal for modeling

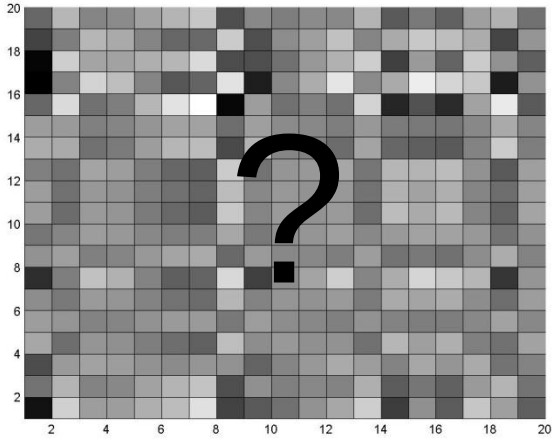
- *Decompose* concentration matrix into **sparse** and **low-rank** components
  - Sparse component for conditional graphical model
  - Low-rank component for latent variables

# Sparse / Low-rank matrix decomposition

$$C = A^* + B^*$$







Given Composite matrix

Unknown Sparse Matrix

Unknown support, values

Unknown Low-rank Matrix

Unknown rank, eigenvectors

**Task: given *C*, recover *A** and *B****

Chandrasekaran *et. al (2009)*

# Applications of sparse/low-rank decompositions

**Statistical model selection**

- Sparse matrix → sparse graphical model
- Low-rank matrix → effect of unobserved latent variables

**Matrix rigidity**

- Change as few entries as possible to make matrix low-rank
- Related to problems in communication complexity

**Composite system identification**

- Sparse matrix → sparse impulse response system
- Low-rank matrix → low model order system

Also, "Robust PCA", face recognition, (Candès-Li-Ma-Wright 2009), etc.

# Sparse/Low-rank decomposition

Let $C = A^* + B^*$, where $A^*$ is sparse and $B^*$ is low-rank.
A possible approach:

$$\min_{A,B} \quad \gamma \, ||A||_0 + rank(B)$$
$$\text{s.t.} \quad A + B = C$$

- combinatorial, NP-hard in general. Cannot solve this efficiently!

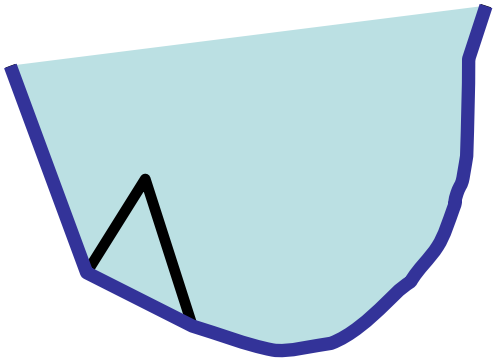
Also,

- not known how to choose $\gamma$
- when does this exactly recover $(A^*, B^*)$ ?

# Convex relaxations

- Bad nonconvex problem -> Convexify!

nonconvex
function



convex
envelope

# Natural convex relaxation

$$||A||_0 \longrightarrow ||A||_1 = \sum_{i,j} |a_{ij}|$$

$$rank(B) = ||\underline{\sigma}(B)||_0 \longrightarrow ||B||_* = \sum_i \sigma_i(B)$$

Propose:

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \quad \gamma\|A\|_1 + \|B\|_*$$

$$\text{s.t.} \quad A + B = C$$

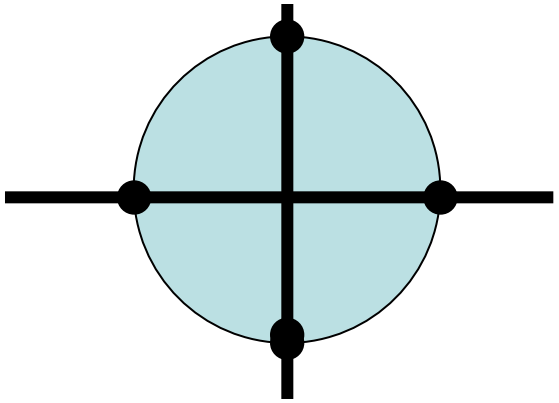
Convex program (in fact, an SDP).

Sufficient conditions for recovery (Chandrasekaran et al. 2009, Candès et al. 2009).

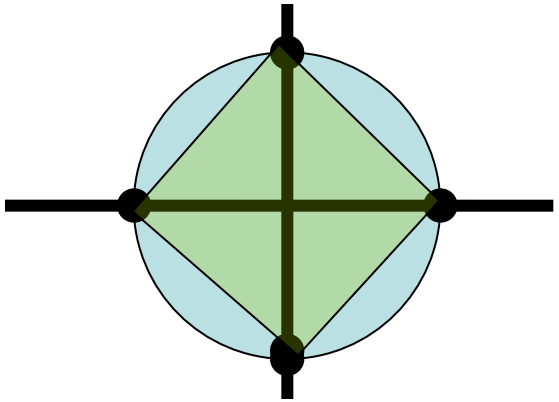
# Inducing sparsity: L1

Consider sparsity minimization

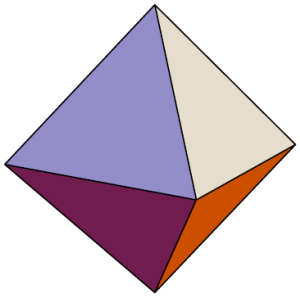
- Geometric interpretation



- Take “sparsity 1” variety
- Intersect with unit ball
- Take convex hull

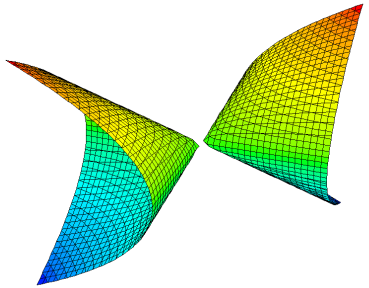


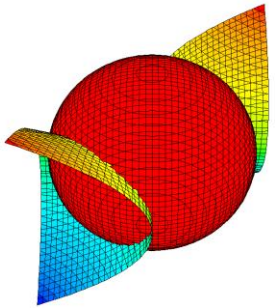
**L1 ball! (crosspolytope)**



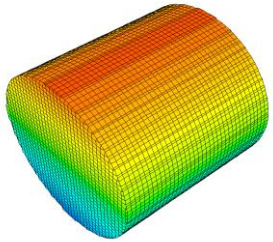
# Inducing low-rank: Nuclear norm

- Same idea!





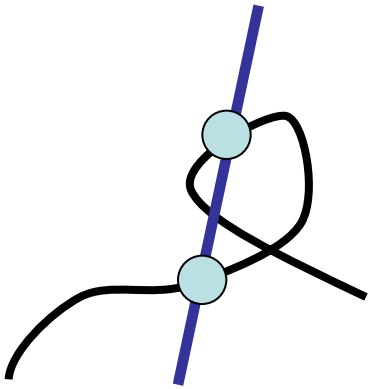
- Take “rank 1” variety
- Intersect with unit ball
- Take convex hull



$$\text{conv}\{uv^T \ : \ u \in \mathbb{R}^n, v \in \mathbb{R}^m, ||u|| = 1, ||v|| = 1\}$$

**Nuclear ball!**

# Beyond rank and sparsity

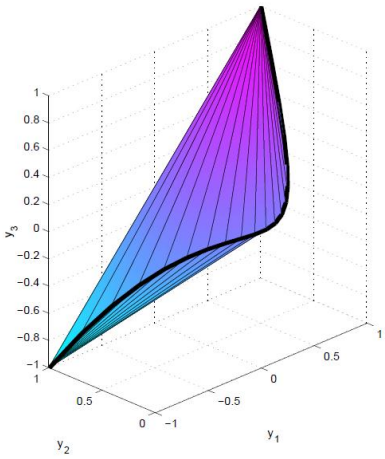


- What is common to these two cases? Can this be further extended?

- Generalize notions of rank to other objects (e.g., tensors, nonnegative matrices, etc.) through *secant varieties* and *atomic norms.*

- Many nice properties (e.g., number of measurements), some technical difficulties (varieties may not be closed, norms may not be polytime computable).

More details in Ben Recht's talk (Wednesday)

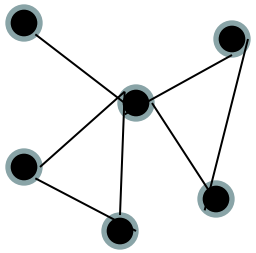


# Proposal for modeling

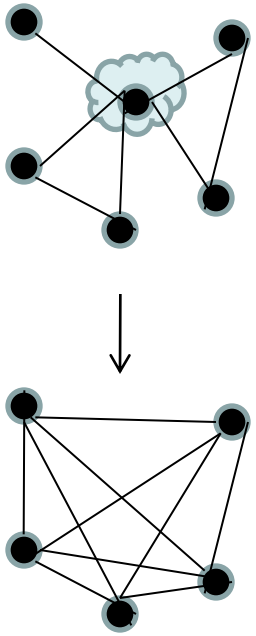
- *Decompose* concentration matrix into **sparse** and **low-rank** components
  - Sparse component for conditional graphical model
  - Low-rank component for latent variables
- Learn sparse graphical model *conditioned* on a few *additional* hidden components
- *Blend* of dimensionality reduction (low-rank) and graphical modeling (sparse)
- Do this in a statistical meaningful way

# Gaussian graphical model framework

- Everything observed



- Some variables not observed
  - Interactions appear *very dense*
  - Graph seems fully connected
  - Sparse modeling not useful



- How to learn a ***simple*** model?

# Covariance estimation via optimization

- Given sample covariance of *n* samples of observed variables: $\Sigma_O^n = \frac{1}{n}\sum_{i=1}^{n} X_O^i (X_O^i)^T$
- Estimate true covariance via *maximum-likelihood*
- Structure via regularization
- For instance, in sparse graphical modelling

$$\hat{S}_n = \arg\min_S \ \mathrm{tr}[S\ \Sigma_O^n] - \log\det(S) \ + \ \lambda_n \|S\|_1$$
$$\text{s.t.} \quad S \succ 0.$$

Banerjee et al. (2006), Ravikumar et al. (2008), …

# Model selection via convex optimization

- Given sample covariance of *n* samples of observed variables:

$$\Sigma_O^n = \frac{1}{n}\sum_{i=1}^{n} X_O^i (X_O^i)^T$$

- Regularized *maximum-likelihood*

Rank/sparsity tradeoff

$$(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} \boxed{\operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L)} + \lambda_n[\gamma\|S\|_1 + \operatorname{tr}(L)]$$

$$\text{s.t.} \quad S - L \succ 0, \quad L \succeq 0.$$

Negative log-likelihood

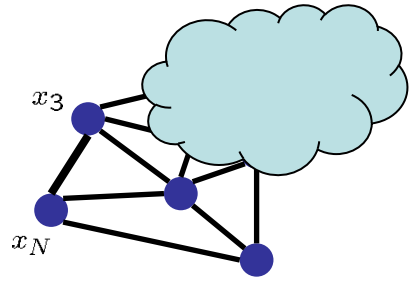
Regularization

# Model selection via convex optimization

$$(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} \boxed{\operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L)} + \lambda_n[\gamma\|S\|_1 + \operatorname{tr}(L)]$$
$$\text{s.t.} \quad S - L \succ 0, \quad L \succeq 0.$$

- Strictly convex optimization
- Typically, large-scale
- Want to understand
  - *consistency* properties (do we get the "right" model?)
  - *sample complexity* (how many samples do we need?)

# Graphical model selection




$$(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} \operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \operatorname{tr}(L)]$$
$$\text{s.t.} \quad S - L \succ 0, \quad L \succeq 0.$$

Under suitable identifiability conditions, and parameters, the estimate given by the convex program yields the correct sign and support for $S_n$, and the correct rank for $L_n$. Explicit rates estimates are available.

Geometric conditions, related to *curvature* of rank varieties.

# Convex Optimization

- Unlike EM-based methods we have
  - *Convex* program
  - *Unique optimum*
  - *Consistency* guarantees

- Parallels with usual sparse graphical modeling

$$\hat{S}_n = \arg\min_{S} \; \mathrm{tr}[S \; \Sigma_O^n] - \log\det(S) \; + \; \lambda_n \|S\|_1$$
$$\text{s.t.} \quad S \succ 0.$$

# Analysis setup

- (S*,L*) true sparse/low-rank components of model from which samples are drawn
  - S* conditional graphical model
  - L* effect of latent vars.
- *p* – # observed vars.
- *n* – # samples
- *h* – # latent vars. (unknown) = rank(L*)
- *High-dimensional* scaling
  - *(p,h,n)* allowed to grow simultaneously

# Assumptions - Identifiability

- $\deg(S^*)$ = max. degree of cond'l graphical model
- $\mathrm{inc}(L^*) = \max_i \ \|P_{\mathrm{rowspace(L^*)}}(e_i)\|$
  - Small value → effect of latent vars. is *spread out* over many observed vars.
- Main condition for identifiability

$$\deg(S^*) \times \mathrm{inc}(L^*) = \mathcal{O}(1)$$

Depends on Fisher information at true model (S*,L*)

# Assumptions – Sample complexity

- # samples $n \gtrsim \frac{p}{\mathrm{inc}(L^*)^4}$
- Min. nonzero entry of S* $\gtrsim \frac{1}{\mathrm{inc}(L^*)\deg(S^*)}\sqrt{\frac{p}{n}}$
- Min. nonzero singular value of L* $\gtrsim \frac{1}{\mathrm{inc}(L^*)^3}\sqrt{\frac{p}{n}}$
- Choose $\gamma \in \left(\frac{6\ \mathrm{inc}(L^*)}{C}, \frac{C}{2\ \deg(S^*)}\right)$

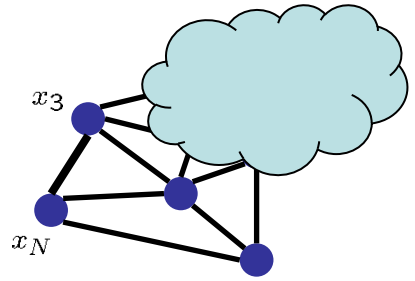
$$\lambda_n \sim \frac{1}{\mathrm{inc}(L^*)}\sqrt{\frac{p}{n}}$$

# High-dimensional consistency

- **Theorem**: Under conditions of previous slides
  - With probability $\geq 1 - \exp\{-C'p\}$
  - Support/sign-pattern of $S^*$ and $\hat{S}_n$ are the same
  - Rank of $L^*$ and $\hat{L}_n$ are the same
  - Error $\sim \lambda_n$ between $(\hat{S}_n, \hat{L}_n)$ and $(S^*, L^*)$

- *Consistently* recover cond'l graphical model of observed vars., and # latent vars.

# Scaling regimes




Let *(p,h)* be the number of observed and latent variables, and *n* the number of samples. Different regimes for coherent estimation:

- Bounded degree:

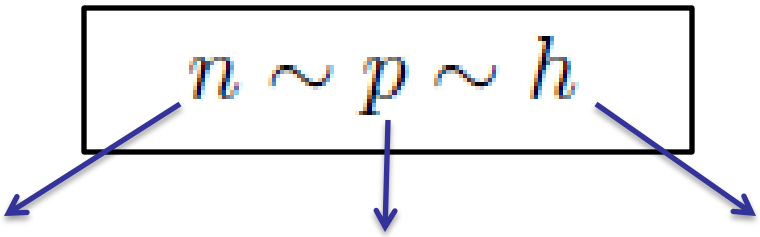
$$d = O(1), \quad h = O(p), \quad n = O(p)$$

- Polylogarithmic degree:

$$d = \mathrm{O}((\log p)^{\mathrm{q}}), \quad h = \mathrm{O}(\mathrm{p}/(\log p)^{2\mathrm{q}}), \quad n = \mathrm{O}(\mathrm{p\ polylog}\ p)$$

# Bounded-degree scaling regime

- Suppose $\mathrm{inc}(L^*) = \mathcal{O}\left(\sqrt{\frac{h}{p}}\right)$
  - rank(L*) = *h*
  - Effect of latent vars. on *most* observed vars.
- Suppose $\deg(S^*) = \mathcal{O}(1)$
  - Cond'l graphical model has *bounded-degree*
- Scaling for consistency:




# samples # obs. vars. # latent vars.

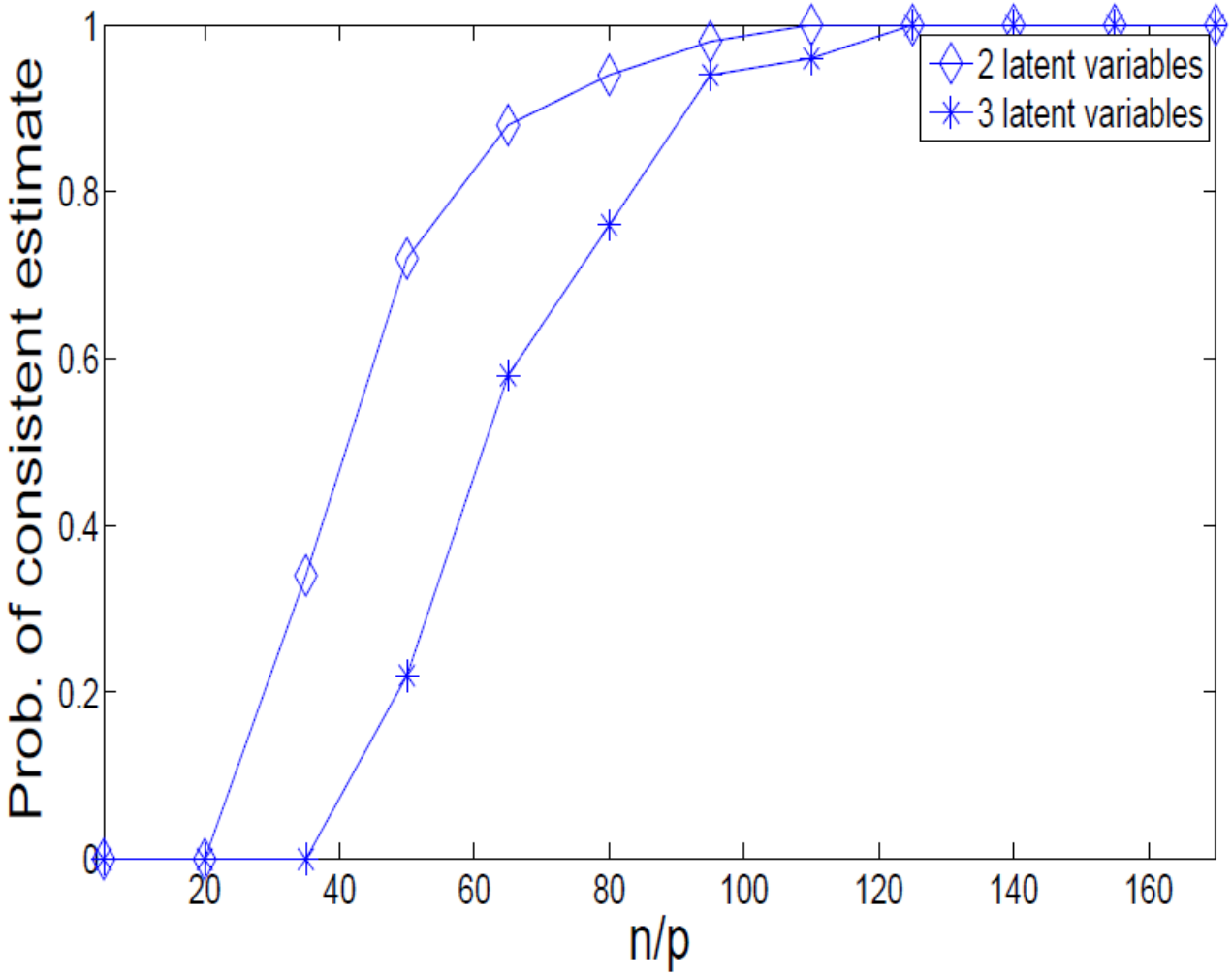
# Algorithms

Convex, nondifferentiable, special structure.

$$(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} \operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \operatorname{tr}(L)]$$
$$\text{s.t.} \quad S - L \succ 0, \quad L \succeq 0.$$

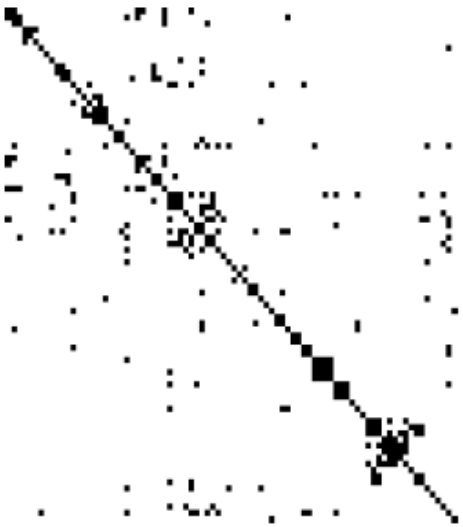
- Possible approaches:
  - Interior-point methods logdet/SDP (e.g., SDPT3)
  - Newton CG primal proximal point algorithm (Wang-Sun-Toh 09). Implemented in **LogDetPPA**
- Adapt others methods from low-rank/sparse opt?
  - Alternating directions (Yuan-Yang)
  - Augmented Lagrangian schemes (Lin et al.)

# Example 1: synthetic data




- 36-cycle among observed variables
- Each hidden variable connected to 80% observed variables

# Example 2: Stock returns



Monthly returns of 84 companies listed in S&P 100

# samples: 216 (1990 to 2007)

Gaussian graphical model without latent variables

**# parameters: 730**
**KL divergence: 44.4**

Gaussian graphical model conditioned on 5 latent variables

**# parameters: 639**
**KL divergence: 17.7**

***Strongest edges***:
AT&T – Verizon
Intel – TI
Apple – Dell

# Thank you!

Want to know more? Details below, and in references therein:

- **V. Chandrasekaran, P.A. Parrilo, A. Willsky, *Latent variable graphical model selection via convex optimization*, arXiv:1008.1290, 2010.**

- V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, A. Willsky, *Rank-Sparsity Incoherence for Matrix Decomposition*, arXiv:0906.2220, 2009.

Thanks for your attention!