School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

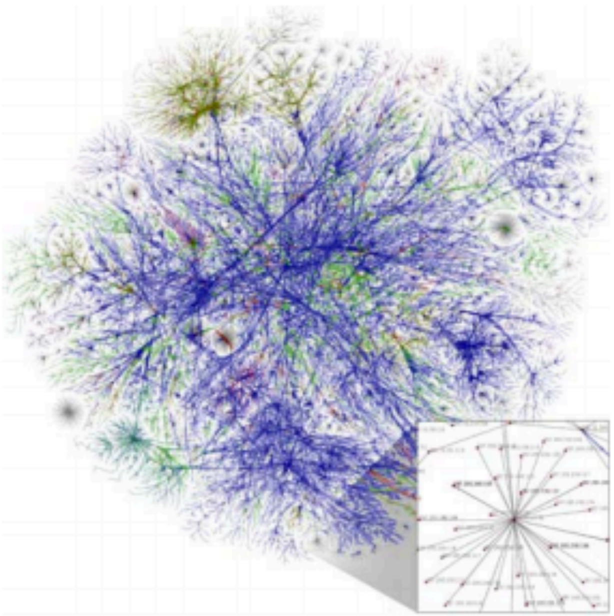
Distributed ADMM for Gaussian Graphical Models

Yaoliang Yu

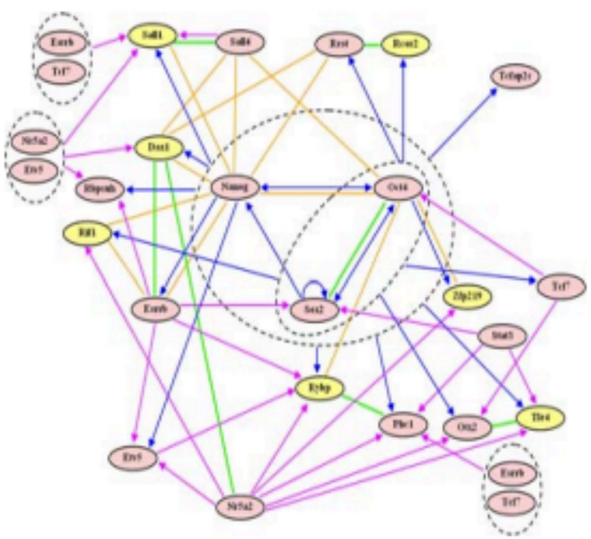
Lecture 29, April 29, 2015

Networks / Graphs

Social Network



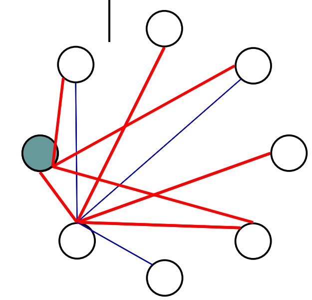
Internet



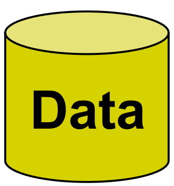
Regulatory Network

Where do graphs come from?

- Prior knowledge
  - Mom told me "A is connected to B"
- Estimate from data!
  - We have seen this in previous classes
  - Will see two more today
- Sometimes may also be interested in edge weights
  - An easier problem
- Real networks are BIG
  - Require distributed optimization



Structural Learning for completely observed MRF (Recall)



$$(x_1^{(1)}, \ldots, x_n^{(1)})$$
$$(x_1^{(2)}, \ldots, x_n^{(2)})$$
$$\ldots$$
$$(x_1^{(M)}, \ldots, x_n^{(M)})$$

Gaussian Graphical Models

- Multivariate Gaussian density:

$$p(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

- WOLG: let $\mu = 0 \quad Q = \Sigma^{-1}$

$$p(x_1, x_2, \cdots, x_p \mid \mu = 0, Q) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} \exp\left\{-\tfrac{1}{2}\sum_i q_{ii}(x_i)^2 - \sum_{i<j} q_{ij} x_i x_j\right\}$$

- We can view this as a continuous Markov Random Field with potentials defined on every node and edge:

The covariance and the precision matrices

- Covariance matrix Σ

$$\Sigma_{i,j} = 0 \quad \Rightarrow \quad X_i \perp X_j \quad \text{or} \quad p(X_i, X_j) = p(X_i)p(X_j)$$

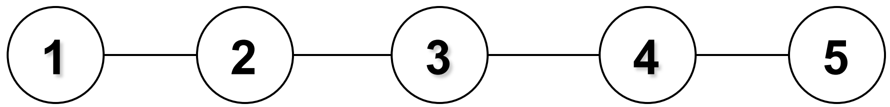
  - Graphical model interpretation?

- Precision matrix $Q = \Sigma^{-1}$

$$Q_{i,j} = 0 \quad \Rightarrow \quad X_i \perp X_j | \mathbf{X}_{-ij} \quad \text{or} \quad p(X_i, X_j | \mathbf{X}_{-ij}) = p(X_i | \mathbf{X}_{-ij}) p(X_j | \mathbf{X}_{-ij})$$

  - Graphical model interpretation?

Sparse precision vs. sparse covariance in GGM



$$\Sigma^{-1} = \begin{pmatrix} 1 & 6 & 0 & 0 & 0 \\ 6 & 2 & 7 & 0 & 0 \\ 0 & 7 & 3 & 8 & 0 \\ 0 & 0 & 8 & 4 & 9 \\ 0 & 0 & 0 & 9 & 5 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 0.10 & 0.15 & -0.13 & -0.08 & 0.15 \\ 0.15 & -0.03 & 0.02 & 0.01 & -0.03 \\ -0.13 & 0.02 & 0.10 & 0.07 & -0.12 \\ -0.08 & 0.01 & 0.07 & -0.04 & 0.07 \\ 0.15 & -0.03 & -0.12 & 0.07 & 0.08 \end{pmatrix}$$

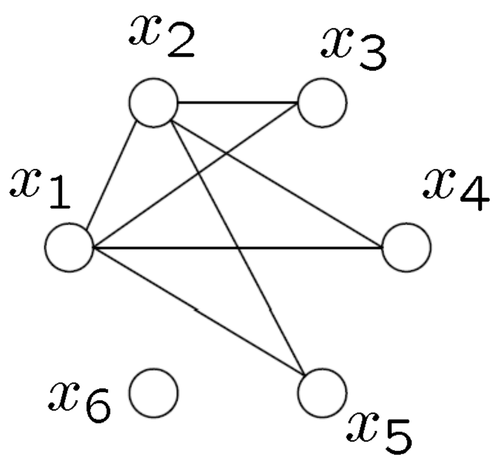
$$\Sigma^{-1}_{15} = 0 \Leftrightarrow X_1 \perp X_5 \,\big|\, X_{nbrs(1) \text{ or } nbrs(5)}$$

$$\not\Rightarrow$$

$$X_1 \perp X_5 \Leftrightarrow \Sigma_{15} = 0$$

Another example

$$Q = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$



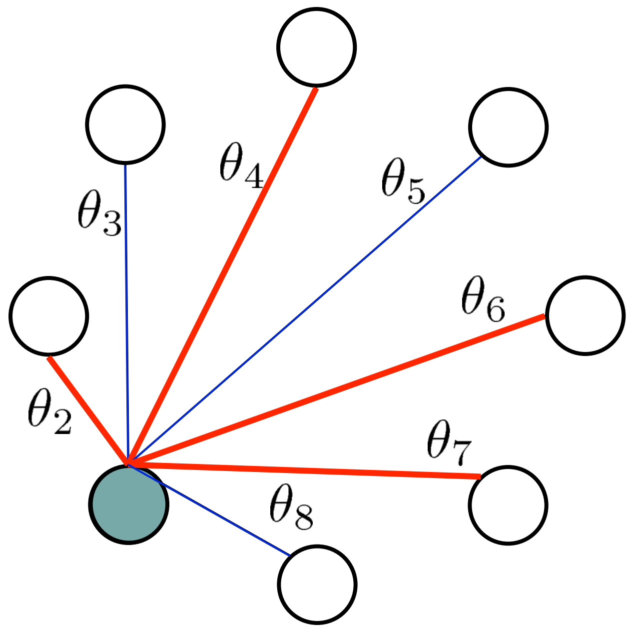
- How to estimate this MRF?
- What if $p >> n$
  - MLE does not exist in general!
  - What about only learning a "sparse" graphical model?
    - This is possible when $s=o(n)$
    - Very often it is the structure of the GM that is more interesting …

Recall lasso

$$\hat{\theta}_i = \arg\min_{\theta_i} l(\theta_i) + \lambda_1 \| \theta_i \|_1$$

$$\text{where } \; l(\theta_i) = \log P\big(y_i|\mathbf{x}_i, \theta_i\big).$$

Graph Regression (Meinshausen & Buhlmann'06)

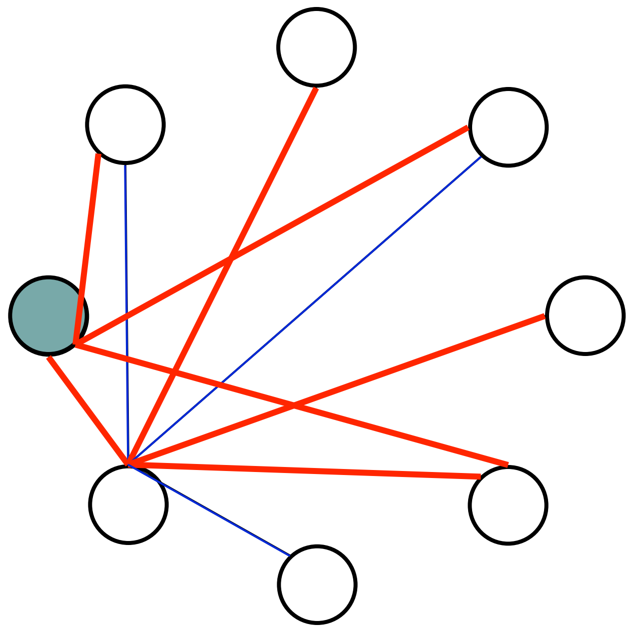


Neighborhood selection

Lasso:

$$\hat{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} l(\theta) + \lambda_1 \| \theta \|_1$$

Graph Regression



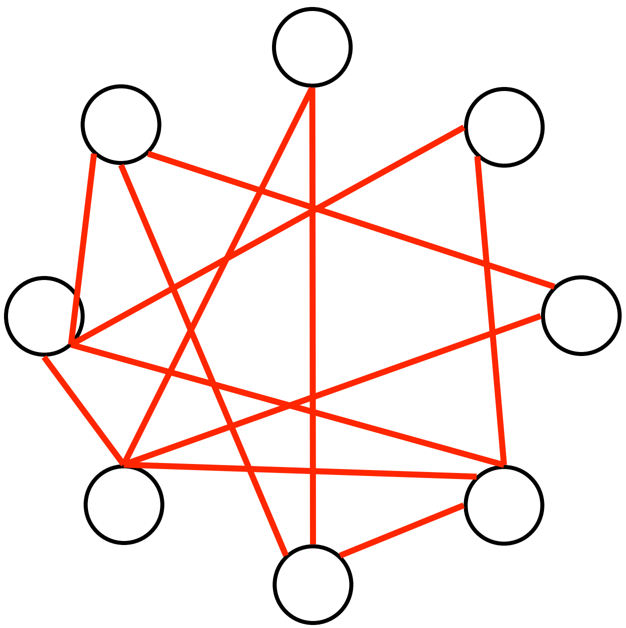
Graph Regression

Pros:

- Computationally convenient
- Strong theoretical guarantee (p <= pol(n))

Cons:

- Asymmetry
- Not minimax optimal



The regularized MLE (Yuan & Lin'07)

$$\min_{Q} -\log\det Q + \mathrm{tr}(QS) + \lambda\|Q\|_1$$

- S: sample covariance matrix, may be singular
- $||Q||_1$: may exclude the diagonal
- log det Q: implicitly force Q to be PSD symmetric

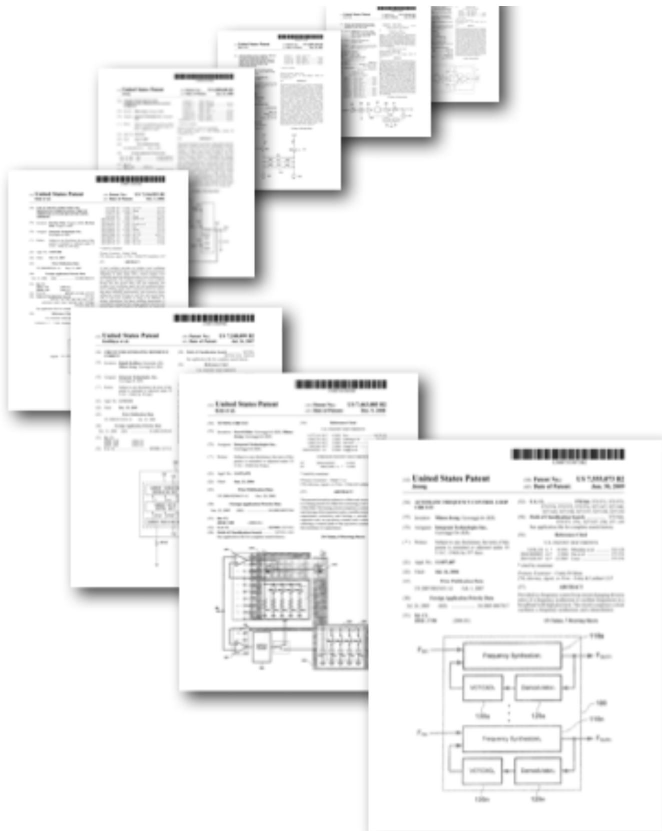
Pros

- Single step for estimating graph and inverse covariance
- MLE!

Cons

- Computationally challenging, partly solved by Glasso (Banergee et al'08, Friedman et al'08)

Many many follow-ups



A closer look of RMLE

$$\min_{Q} -\log\det Q + \mathrm{tr}(QS) + \lambda\|Q\|_1$$

- Set derivative to 0:

$$-Q^{-1} + S + \lambda \cdot \mathrm{sign}(Q) = 0$$

$$\Downarrow$$

$$\|Q^{-1} - S\|_\infty \leq \lambda$$

- Can we (?!):

$$\min_{Q} \|Q\|_1 \text{ s.t. } \|Q^{-1} - S\|_\infty \leq \lambda$$

CLIME (Cai et al.'11)

- Further relaxation

$$\min_{Q} \|Q\|_1 \text{ s.t. } \|SQ - I\|_\infty \leq \lambda$$

  - Constraint controls $Q \approx S^{-1}$
  - Objective controls sparsity in Q
  - Q is not required to be PSD or symmetric

- Separable! LP!!!
  - Both objective and constraint are element-wise separable
  - Can be reformulated as LP

- Strong theoretical guarantee
  - Variations are minimax-optimal (Cai et al.'12, Liu & Wang'12)

But for BIG problems

$$\min_{Q} \|Q\|_1 \text{ s.t. } \|SQ - I\|_\infty \leq \lambda$$

- Standard solvers for LP can be slow
- Embarrassingly parallel:
  - Solve each column of Q independently in each core/machine

$$\min_{q_i} \|q_i\|_1 \text{ s.t. } \|Sq_i - e_i\|_\infty \leq \lambda$$

  - Thanks for not having PSD constraint on Q
  - Still troublesome if S is big
  - Need to consider first-order methods

A gentle introduction to alternating direction method of multipliers (ADMM)

Optimization with coupling variables

$$\text{Canonical form: } \min_{w,z} \overbrace{f(w) + g(z)}^{\text{☺ uncoupled}}, \quad \text{s.t.} \quad \overbrace{Aw + Bz = c}^{\text{☹ coupled}},$$

where $w \in \mathbb{R}^m, z \in \mathbb{R}^p, A : \mathbb{R}^m \to \mathbb{R}^q, B : \mathbb{R}^p \to \mathbb{R}^q, c \in \mathbb{R}^q$

- Numerically challenging because
  - Function f or g nonsmooth or constrained (i.e., can take value ∞)
  - Linear constraint couples the variables w and z
  - Large scale, interior point methods NA
- Naively alternating x and z does not work
  - Min w^2 s.t. $w + z = 1$; optimum clearly is $w = 0$
  - Start with say $w = 1 \to z = 0 \to w = 1 \to z = 0 \ldots$
- However, without coupling, can solve separately w and z
  - Idea: try to decouple vars in the constraint!

Example: Empirical Risk Minimization (ERM)

$$\min_w g(w) + \overbrace{\sum_{i=1}^{n} f_i(w)}^{\text{☹ coupled}}$$

- Each i corresponds to a training point (x_i, y_i)
- Loss f_i measures the fitness of the model parameter w
  - least squares: $f_i(w) = (y_i - w^\top x_i)^2$
  - support vector machines: $f_i(w) = (1 - y_i w^\top x_i)_+$
  - boosting: $f_i(w) = \exp(-y_i w^\top x_i)$
  - logistic regression: $f_i(w) = \log(1 + \exp(-y_i w^\top x_i))$
- g is the regularization function, e.g. $\lambda_n \|w\|_2^2$ or $\lambda_n \|w\|_1$
- Vars coupled in obj, but not in constraint (none)
  - Reformulate: transfer coupling from obj to constraint
  - Arrive at canonical form, allow unified treatment later

Why canonical form?

ERM: $$\min_{w} g(w) + \sum_{i=1}^{n} f_i(w)$$

Canonical form: $$\min_{w,z} f(w) + g(z), \quad \text{s.t.} \quad Aw + Bz = c,$$

where $w \in \mathbb{R}^m, z \in \mathbb{R}^p, A : \mathbb{R}^m \to \mathbb{R}^q, B : \mathbb{R}^p \to \mathbb{R}^q, c \in \mathbb{R}^q$

- ADMM algorithm (to be introduced shortly) excels at solving the canonical form
  - Canonical form is a general "template" for constrained problems
- ERM (and many other problems) can be converted to canonical form through variable duplication (see next slide)

How to: variable duplication

- Duplicate variables to achieve canonical form

$$\min_{w} g(w) + \sum_{i=1}^{n} f_i(w)$$

$$\Big\Downarrow \quad v = [w_1, \ldots, w_n]^\top$$

$$\min_{v,z} \quad g(z) + \underbrace{\sum\nolimits_i f_i(w_i)}_{f(v)}, \quad \text{s.t.} \quad \underbrace{w_i = z, \forall i}_{v-[I,\ldots,I]^\top z=0}$$

- Global consensus constraint: $\forall i, w_i = z$
  - All w_i must (eventually) agree
- Downside: many extra variables, increase problem size
  - Implicitly maintain duplicated variables

Augmented Lagrangian

Canonical form: $$\min_{\mathbf{w},\mathbf{z}} f(\mathbf{w}) + g(\mathbf{z}), \quad \text{s.t.} \quad A\mathbf{w} + B\mathbf{z} = \mathbf{c},$$

where $\mathbf{w} \in \mathbb{R}^m, \mathbf{z} \in \mathbb{R}^p, A : \mathbb{R}^m \to \mathbb{R}^q, B : \mathbb{R}^p \to \mathbb{R}^q, \mathbf{c} \in \mathbb{R}^q$

- Intro Lagrangian multiplier $\boldsymbol{\lambda}$ to decouple variables

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \underbrace{f(\mathbf{w}) + g(\mathbf{z}) + \boldsymbol{\lambda}^\top (A\mathbf{w} + B\mathbf{z} - \mathbf{c}) + \frac{\mu}{2}\|A\mathbf{w} + B\mathbf{z} - \mathbf{c}\|_2^2}_{L_\mu(\mathbf{w},\mathbf{z};\boldsymbol{\lambda})}$$

  - L_μ : augmented Lagrangian
  - More complicated min-max problem, but no coupling constraints

Why Augmented Lagrangian?

- Quadratic term gives numerical stability

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \underbrace{f(\mathbf{w}) + g(\mathbf{z}) + \boldsymbol{\lambda}^\top (A\mathbf{w} + B\mathbf{z} - \mathbf{c}) + \boxed{\frac{\mu}{2} \|A\mathbf{w} + B\mathbf{z} - \mathbf{c}\|_2^2}}_{L_\mu(\mathbf{w},\mathbf{z};\boldsymbol{\lambda})}$$

- May lead to strong convexity in w or z
  - Faster convergence when strongly convex
- Allows larger step size (due to higher stability)
- Prevents subproblems diverging to infinity (again, stability)

- But sometimes better to work with normal Lagrangian

ADMM Algorithm

$$\min_{\mathbf{w},\mathbf{z}} \max_{\boldsymbol{\lambda}} \underbrace{f(\mathbf{w}) + g(\mathbf{z}) + \boldsymbol{\lambda}^\top (A\mathbf{w} + B\mathbf{z} - \mathbf{c}) + \frac{\mu}{2}\|A\mathbf{w} + B\mathbf{z} - \mathbf{c}\|_2^2}_{L_\mu(\mathbf{w},\mathbf{z};\boldsymbol{\lambda})}$$

- Fix dual $\boldsymbol{\lambda}$, block coordinate descent on primal w, z

$$\mathbf{w}^{t+1} \leftarrow \arg\min_{\mathbf{w}} L_\mu(\mathbf{w}, \mathbf{z}^t; \boldsymbol{\lambda}^t) \equiv f(\mathbf{w}) + \frac{\mu}{2}\|A\mathbf{w} + B\mathbf{z}^t - \mathbf{c} + \boldsymbol{\lambda}^t/\mu\|^2$$

$$\mathbf{z}^{t+1} \leftarrow \arg\min_{\mathbf{z}} L_\mu(\mathbf{w}^{t+1}, \mathbf{z}; \boldsymbol{\lambda}^t) \equiv g(\mathbf{z}) + \frac{\mu}{2}\|A\mathbf{w}^{t+1} + B\mathbf{z} - \mathbf{c} + \boldsymbol{\lambda}^t/\mu\|^2$$

- Fix primal w, z, gradient ascent on dual $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda}^{t+1} \leftarrow \boldsymbol{\lambda}^t + \eta(A\mathbf{w}^{t+1} + B\mathbf{z}^{t+1} - \mathbf{c})$$

- Step size η can be large, e.g. $\eta = \mu$
  - Usually rescale $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}/\eta$ to remove η

ERM revisited

- Reformulate by duplicating variables

$$\min_{\mathbf{v},\mathbf{z}} \quad g(\mathbf{z}) + \underbrace{\sum_i f_i(\mathbf{w}_i)}_{f(\mathbf{v})}, \quad \text{s.t.} \quad \underbrace{\mathbf{w}_i = \mathbf{z}, \forall i}_{\mathbf{v}-[I,\ldots,I]^\top \mathbf{z}=0}$$

- ADMM x-step:

$$\mathbf{w}^{t+1} \leftarrow \arg\min_{\mathbf{w}} L_\mu(\mathbf{w}, \mathbf{z}^t; \boldsymbol{\lambda}^t) \equiv f(\mathbf{w}) + \frac{\mu}{2}\|A\mathbf{w} + B\mathbf{z}^t - \mathbf{c} + \boldsymbol{\lambda}^t/\mu\|^2$$

$$= \sum_i \underbrace{f_i(\mathbf{w}_i) + \frac{\mu}{2}\|\mathbf{w}_i - \mathbf{z}^t + \boldsymbol{\lambda}_i^t\|^2}$$

- Thanks to duplicating

  - **Completely decoupled**
  - **Parallelizable**
  - **Closed-form if f_i is "simple"**

ADMM: History and Related

- Augmented Lagrangian Method (ALM): solve w, z jointly even though coupled
  - (Bertsekas'82) and refs therein
- Alternating Direction of Multiplier Method (ADMM): alternate w and z as previous slide
  - (Boyd et al.'10) and refs therein
  - Operator splitting for PDEs: Douglas, Peaceman, and Rachford (50s-70s)
  - Glowinsky et al.'80s, Gabay'83; Spingarn'85
  - (Eckstein & Bertsekas'92; He et al.'02) in variational inequality
  - Lots of recent work.

ADMM: Linearization

- Demanding step in each iteration of ADMM (similar for z):

$$x_{t+1} \leftarrow \arg\min_x L_\mu(x, z_t; y_t) = f(x) + g(z_t) + y_t^\top (Ax + Bz_t - c) + \frac{\mu}{2}\|Ax + Bz_t - c\|_2^2$$

- Diagonal A: reduce to proximal map (more later) of f
  - $f(x) = \|x\|_1$, soft-shrinkage: $\mathrm{sign}(x) \cdot (|x| - \mu)_+$
- Non-diagonal A: no closed-form, messy inner loop
- Instead, reduce to diagonal A by
  - A single gradient step: $x_{t+1} \leftarrow x_t - \eta \partial f(x_t) + A^\top y_t + \mu A^\top (Ax_t + Bz_t - c)$
  - Or, linearize the quadratic at x_t:

$$x_{t+1} \leftarrow \arg\min_x \underbrace{f(x) + y_t^\top Ax + (x - x_t)^\top \mu A^\top (Ax_t + Bz_t - c) + \frac{\mu}{2}\|x - x_t\|_2^2}_{f(x) + \frac{\mu}{2}\|x - x_t + A^\top (Ax_t + Bz_t - c + y_t/\mu)\|_2^2}$$

- Intuition: x re-computed in the next iteration anyways
  - No need for “perfect” x

Convergence Guarantees: Fixed-point theory

- Recall some definitions

$$\text{proximal map } \mathsf{P}_f^\mu(w) := \arg\min_z \tfrac{1}{2\mu}\|z - w\|_2^2 + f(z)$$

$$\text{reflection map } \mathsf{R}_f^\mu(w) := 2\mathsf{P}_f^\mu(w) - w$$

  - well-defined for convex f, non-expansive: $\|T(x) - T(y)\|_2 \leq \|x - y\|_2$
  - proximal map generalizes the Euclidean projection
- Lagrangian: $L_0(x, z; y) = \underbrace{\min_x \Big(f(x) + y^\top Ax\Big)}_{d_1(y)} + \underbrace{\min_z \Big(g(z) + y^\top (Bz - c)\Big)}_{d_2(y)}$
- ADMM = Douglas-Rachford splitting

$$w \leftarrow \tfrac{1}{2}(w + \mathsf{R}_{d_2}^\mu(\mathsf{R}_{d_1}^\mu(w))); \quad y \leftarrow \mathsf{P}_{d_2}^\mu(w)$$

- Fixed-point iteration!
  - convergence follows, e.g. (Bauschke & Combettes'13)
  - explains why dual y, not primal x or z, always converges

ADMM for CLIME

Apply ADMM to CLIME

$$\min_{Q} \|Q\|_1 \text{ s.t. } \|SQ - E\|_\infty \leq \lambda$$

- Solve a block of columns of Q in each core/machine
  - E is the corresponding block in I
- Step 1: reduce to ADMM canonical form
  - Use variable duplicating

$$\min_{Q,Z} \|Q\|_1 \text{ s.t. } \|Z - E\|_\infty \leq \lambda, \quad Z = SQ$$

$$\min_{Q,Z} \|Q\|_1 + [\|Z - E\|_\infty \leq \lambda] \text{ s.t. } Z = SQ$$

Apply ADMM to CLIME (cont')

- Step 2: Write out augmented Lagrangian

$$L(Q,Z;Y) = \|Q\|_1 + [\|Z-E\|_\infty \le \lambda] + \rho \mathrm{tr}[(SQ-Z)Y] + \frac{\rho}{2}\|SQ-Z\|_F^2$$

- Step 3: Perform primal-dual updates

$$Q \leftarrow \arg\min_Q \|Q\|_1 + \frac{\rho}{2}\|SQ-Z+Y\|_F^2$$

$$Z \leftarrow \arg\min_Z [\|Z-E\|_\infty \le \lambda] + \frac{\rho}{2}\|SQ-Z+Y\|_F^2$$

$$= \arg\min_{\|Z-E\|_\infty \le \lambda} \frac{\rho}{2}\|SQ-Z+Y\|_F^2$$

$$Y \leftarrow Y + SQ - Z$$

Apply ADMM to CLIME (cont'')

- Step 4: Solve the subproblems
  - Lagrangian dual Y: trivial
  - Primal Z: projection to l_inf ball, separable, easy
  - Primal Q: easy if S is orthogonal, in general a lasso problem
- Bypass double loop by linearization
  - Intuition: wasteful to solve Q to death

$$\min_Q \|Q\|_1 + \rho \mathrm{tr}(Q^\top S(Y + SQ_t - Z)) + \frac{\eta}{2}\|Q - Q_t\|_F^2$$

  - Soft-thresholding
- Putting things together

$$Q \leftarrow \arg\min_Q \|Q\|_1 + \frac{\rho}{2}\|SQ - Z + Y\|_F^2$$

$$Z \leftarrow \arg\min_Z [\|Z - E\|_\infty \leq \lambda] + \frac{\rho}{2}\|SQ - Z + Y\|_F^2$$

$$= \arg\min_{\|Z-E\|_\infty \leq \lambda} \frac{\rho}{2}\|SQ - Z + Y\|_F^2$$

$$Y \leftarrow Y + SQ - Z$$

Exploring structure

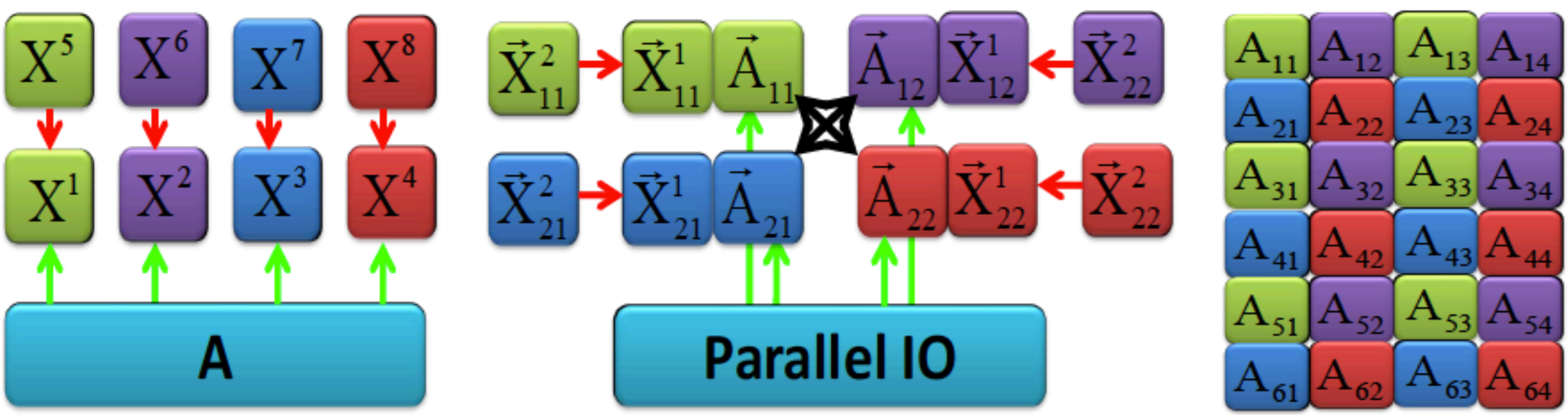
- Expensive step in ADMM-CLIME:
  - Matrix-matrix multiplication: SQ and alike

- If p >> n, S is size p x p but of rank at most n
  - Write S = AA', and do A(A'Q) $A = [X_1, \ldots, X_n] \in \mathbb{R}^{p \times n}$

- Matrix * matrix >> for loop of matrix * vector
  - Preferable to solve a balanced block of columns

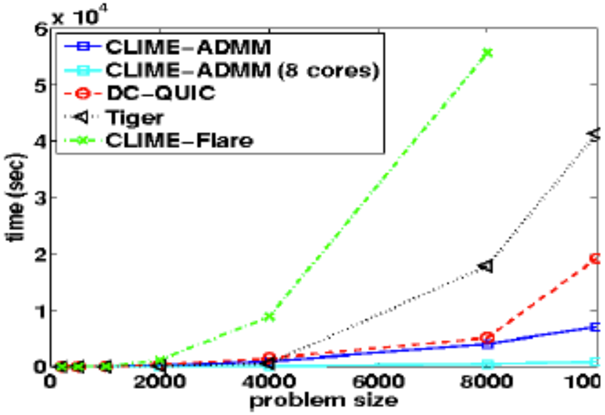
Parallelization (Wang et al'13)

- Embarrassingly parallel
  - If A fits into memory
- Chop A into small blocks and distribute
  - Communication may be high

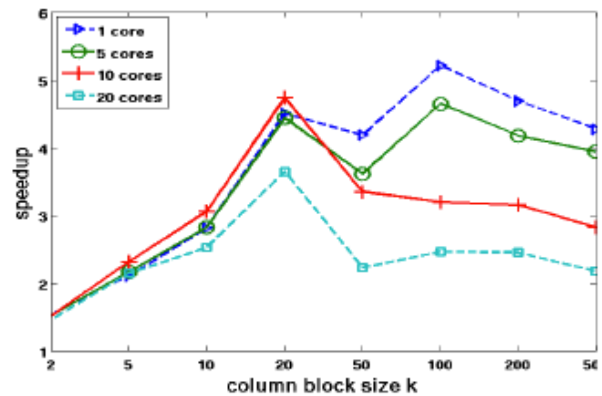


(a) Shared-Memory (b) Distributed-Memory (c) Block Cyclic

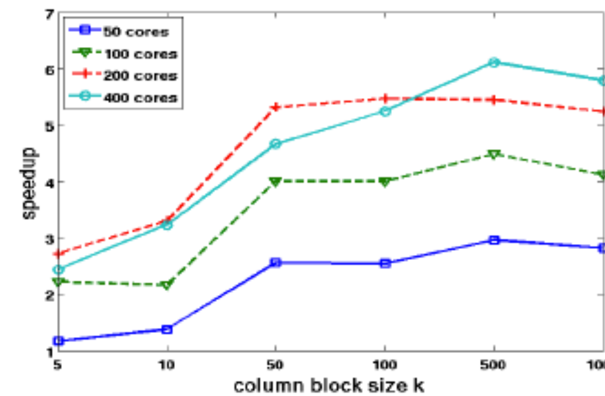
Numerical results (Wang et al'13)



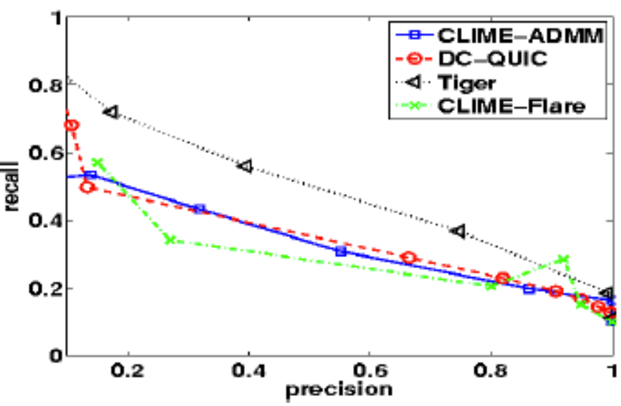
(a) Runtime



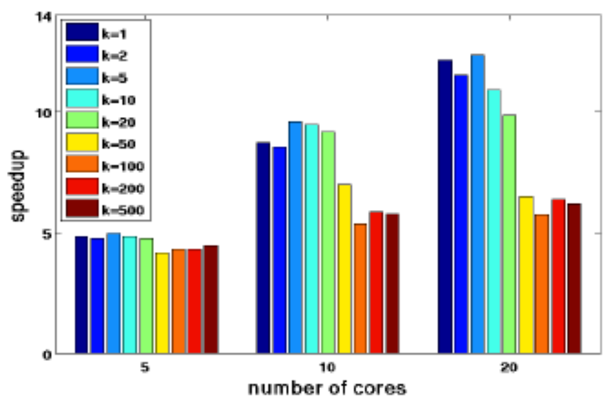
(a) Speedup S_k^{col}



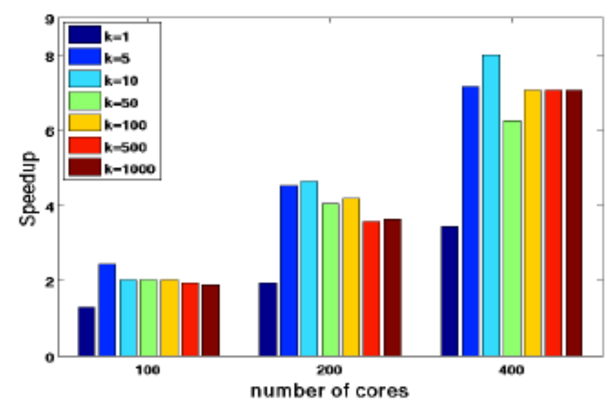
(a) Speedup S_k^{col}



(b) Precision and recall



(b) Speedup S_q^{core}



(b) Speedup S_q^{core}

Numerical results cont' (Wang et al'13)

node ×core	k = 1	k = 5	k = 10	k = 50	k = 100	k = 500	k = 1000
100×1	0.56	1.26	2.59	6.98	13.97	62.35	136.96
25× 4	1.02	2.40	3.42	8.25	16.44	84.08	180.89
200×1	0.37	0.68	1.12	3.48	6.76	33.95	70.59
50×4	0.74	1.44	2.33	4.49	8.33	48.20	103.87

Nonparanormal extensions

Nonparanormal (Liu et al.'09)

$$Z_i = f_i(X_i), \quad i = 1, \ldots, p$$

$$(Z_1, \ldots, Z_p) \sim \mathcal{N}(0, \Sigma)$$

- $f_{i:}$ unknown monotonic functions
- Observe X, but not Z
- Independence preserved under transformations

$$X_i \perp X_j | X_{rest} \iff Z_i \perp Z_j | Z_{rest} \iff Q_{ij} = 0$$

- Can estimate f_i first, then apply glasso on $f_i(X_i)$
  - Estimating functions can be slow, nonparametric rate

A neat observation

- Since f_i is monotonic
  - $Z_{i,:}$ comonotone / concordant with $X_{i,:}$
- Use rank estimator !

$$Z_i = f_i(X_i), \quad i = 1, \ldots, p$$

$$(Z_1, \ldots, Z_p) \sim \mathcal{N}(0, \Sigma)$$

$$Z_{i,1}, Z_{i,2}, \ldots, Z_{i,n}$$

$$\Downarrow$$

$$R_{i,1}, R_{i,2}, \ldots, R_{i,n}$$

$$\Uparrow$$

$$X_{i,1}, X_{i,2}, \ldots, X_{i,n}$$

- Want Σ , but do not observe Z
  - Maybe ???

$$\Sigma_{ij} = \frac{1}{n} \sum_{k=1}^{n} Z_{i,k} Z_{j,k} \stackrel{?}{=} \frac{1}{n} \sum_{k=1}^{n} R_{i,k} R_{j,k}$$

Kendall's tau

- Assuming no ties:

$$\tau_{ij} = \frac{1}{n(n-1)} \sum_{k,\ell} \mathrm{sign}[(R_{i,k} - R_{i,\ell})(R_{j,k} - R_{j,\ell})]$$

  - Complexity of computing Kendall's tau?

- Key: $$\Sigma_{ij} = 2\sin(\frac{\pi}{6}\mathbb{E}(\tau_{ij}))$$

  - Genuine, asymptotically unbiased estimate of Σ
  - $t \mapsto 2\sin(\frac{\pi}{6}t)$ is a contraction, preserving concentration

- After having Σ, use whatever glasso, e.g., CLIME
  - Can also use other rank estimator, e.g., Spearman's rho

Summary

- Gaussian graphical model selection
  - Neighborhood selection, Regularized MLE, CLIME
  - Implicit PSD vs. Explicit PSD
- Distributed ADMM
  - Generic procedure
  - Can distribute matrix product
- Nonparanormal Gaussian graphical model
  - Rank statistics
  - Plug-in estimator

References

- Graph Lasso
  - High-Dimensional Graphs and Variable Selection with the LASSO, Meinshausen and Buhlmann, Annals of Statistics, vol. 34, 2006
  - Model Selection and Estimation in the Gaussian Graphical Model, Yuan and Lin, Biometrika, vol. 94, 2007
  - A Constrained l_1 minimization Approach for Sparse Precision Matrix Estimation, Cai et al., Journal of the American Statistical Association, 2011
  - Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data, Banerjee et al, Journal of Machine Learning Research, 2008
  - Sparse Inverse Covariance Estimation with the Graphical Lasso, Friedman et al., Biostatistics, 2008
- Nonparanormal
  - The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs, Han Liu et al., Journal of Machine Learning Research, 2009
  - Regularized Rank-based Estimation of High-Dimensional Nonparanormal Graphical Models, Lingzhou Xue and Hui Zou, Annals of Statistics, vol. 40, 2012
  - High-Dimensional Semiparametric Gaussian Copula Graphical Models, Han Liu et al., Annals of Statistics, vol. 40, 2012
- ADMM
  - Large Scale Distributed Sparse Precision Estimation, Wang et al., NIPS 2013
  - Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Boyd et al., Foundation and Trends in Machine Learning, 2011