Large-Scale Optimization Algorithms for Sparse Conditional Gaussian Graphical Models

Carnegie Mellon

Calvin McCarter & Seyoung Kim

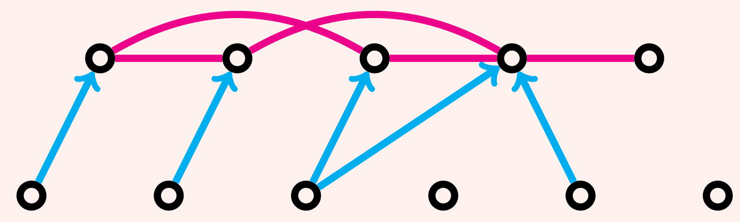
MACHINE LEARNING DEPARTMENT

Motivation

Sparse Conditional Gaussian Graphical Model

$$p(\mathbf{y}|\mathbf{x};\mathbf{\Lambda},\mathbf{\Theta}) = \exp\{-\mathbf{y}^T\mathbf{\Lambda}\mathbf{y} - 2\mathbf{x}^T\mathbf{\Theta}\mathbf{y}\}/Z(\mathbf{x})$$

$$Z(\mathbf{x}) = (2\pi)^{q/2}|\mathbf{\Lambda}|^{-1}\exp(\mathbf{x}^T\mathbf{\Theta}\mathbf{\Lambda}^{-1}\mathbf{\Theta}^T\mathbf{x})$$



Sparse Estimation

Given empirical covariances: $\mathbf{S_{xx}} \in \mathbb{R}^{p\times p}, \mathbf{S_{xy}} = \in \mathbb{R}^{p\times q}, \mathbf{S_{yy}} = \in \mathbb{R}^{q\times q}$

$$\min_{\mathbf{\Lambda}\succ 0,\mathbf{\Theta}} f(\mathbf{\Lambda},\mathbf{\Theta}) = g(\mathbf{\Lambda},\mathbf{\Theta}) + h(\mathbf{\Lambda},\mathbf{\Theta})$$

$$g(\mathbf{\Lambda},\mathbf{\Theta}) = -\log|\mathbf{\Lambda}| + \mathrm{tr}(\mathbf{S_{yy}}\mathbf{\Lambda} + 2\mathbf{S_{xy}}^T\mathbf{\Theta} + \mathbf{\Lambda}^{-1}\mathbf{\Theta}^T\mathbf{S_{xx}}\mathbf{\Theta})$$

$$h(\mathbf{\Lambda},\mathbf{\Theta}) = \lambda_{\mathbf{\Lambda}}\|\mathbf{\Lambda}\|_1 + \lambda_{\mathbf{\Theta}}\|\mathbf{\Theta}\|_1$$

Convex but difficult problem due to last term

Previous Optimization Algorithms

- OWL-QN [1]
- FISTA [2]
- **Newton Coordinate Descent [3]**
  - Second order approximation minimized over active set
  - Proximal Newton subproblem solved via coordinate descent
  - Step size found via backtracking

Second order approximation on both **Λ** and **Θ**

$$\bar{g}_{\mathbf{\Lambda},\mathbf{\Theta}}(\Delta_{\mathbf{\Lambda}},\Delta_{\mathbf{\Theta}}) = \mathrm{vec}(\nabla g(\mathbf{\Lambda},\mathbf{\Theta}))^T\mathrm{vec}([\Delta_{\mathbf{\Lambda}}\ \Delta_{\mathbf{\Theta}}]) + \frac{1}{2}\mathrm{vec}([\Delta_{\mathbf{\Lambda}}\ \Delta_{\mathbf{\Theta}}])^T\nabla^2 g(\mathbf{\Lambda},\mathbf{\Theta})\,\mathrm{vec}([\Delta_{\mathbf{\Lambda}}\ \Delta_{\mathbf{\Theta}}])$$

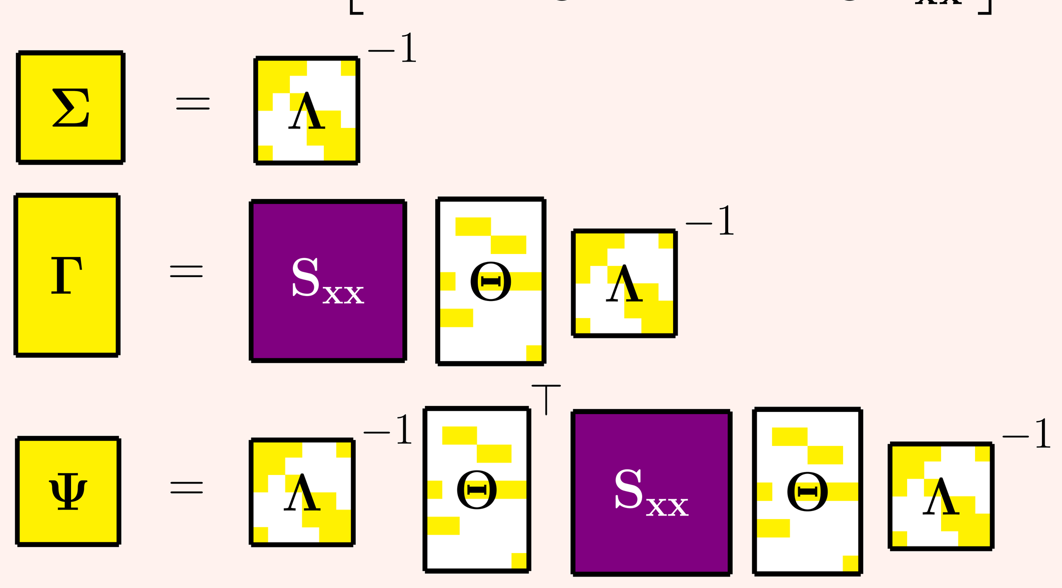
Scalability Problems

Time: Genomic dataset with p = 34k, q = 10k: **> 50 hours**

Memory: Requires $O(pq+q^2)$ memory: **>100 Gb** when p+q = 80k

$$\nabla g(\mathbf{\Lambda},\mathbf{\Theta}) = [\mathbf{S_{yy}} - \mathbf{\Sigma} - \mathbf{\Psi}\quad 2\mathbf{S_{xy}} + 2\mathbf{\Gamma}]$$

$$\nabla^2 g(\mathbf{\Lambda},\mathbf{\Theta}) = \begin{bmatrix}\mathbf{\Sigma}\otimes(\mathbf{\Sigma}+2\mathbf{\Psi}) & -2\mathbf{\Sigma}\otimes\mathbf{\Gamma}^T \\ -2\mathbf{\Sigma}\otimes\mathbf{\Gamma} & 2\mathbf{\Sigma}\otimes\mathbf{S_{xx}}\end{bmatrix}$$



Alternating Newton Coordinate Descent

faster computation

Update Λ given fixed Θ:

- Solve for Newton direction via CD

$$\bar{g}_{\mathbf{\Lambda},\mathbf{\Theta}}(\Delta_{\mathbf{\Lambda}}) = \mathrm{vec}(\nabla_{\mathbf{\Lambda}} g(\mathbf{\Lambda},\mathbf{\Theta}))^T\mathrm{vec}(\Delta_{\mathbf{\Lambda}}) + \tfrac{1}{2}\mathrm{vec}(\Delta_{\mathbf{\Lambda}})^T\nabla^2_{\mathbf{\Lambda}} g(\mathbf{\Lambda},\mathbf{\Theta})\,\mathrm{vec}(\Delta_{\mathbf{\Lambda}})$$

- Run backtracking line search

Update Θ given fixed Λ:

- Solve *Lasso* problem directly via CD

$$g_{\mathbf{\Lambda}}(\mathbf{\Theta}) = \mathrm{tr}(2\mathbf{S_{xy}}^T\mathbf{\Theta} + \mathbf{\Lambda}^{-1}\mathbf{\Theta}^T\mathbf{S_{xx}}\mathbf{\Theta})$$

Second order approximation only on **Λ**

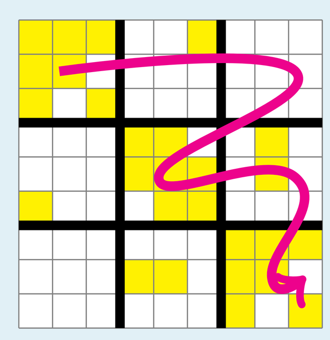
- ✓ Eliminate **Γ**
- ✓ Reduce CD time complexity
- ✓ Backtrack only for **Λ** – faster early convergence

Alternating Newton Block Coordinate Descent

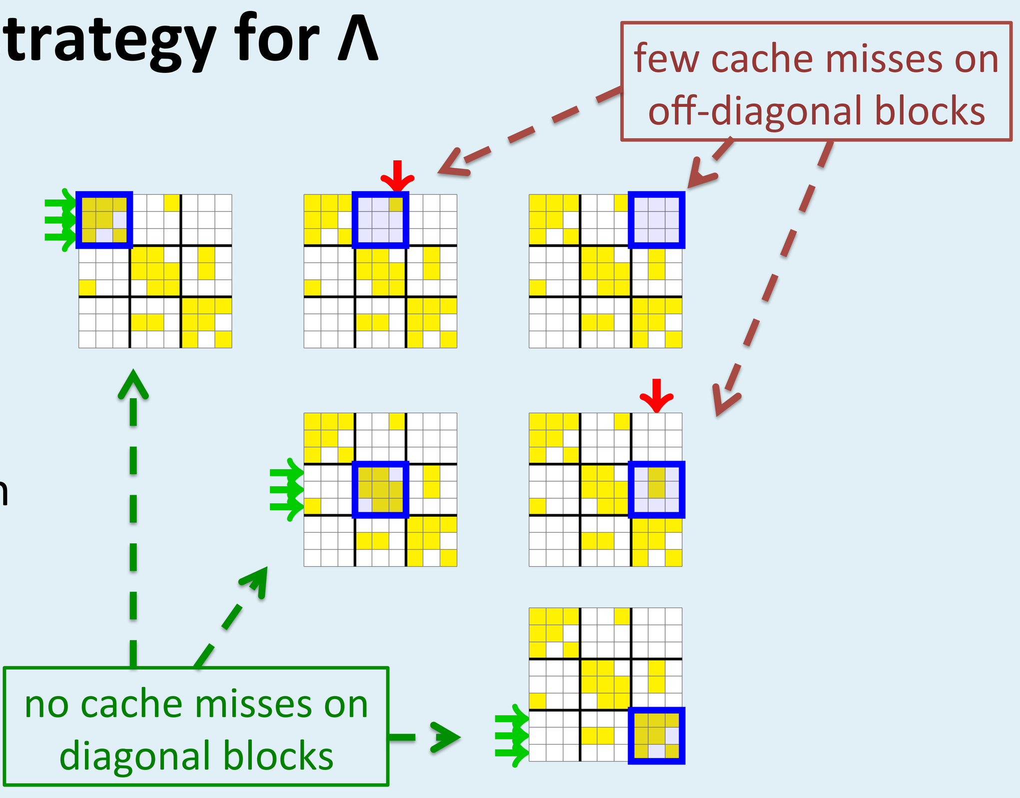
no memory restriction

- ☒ Pre-compute all matrices: lots of memory, fast
- ☒ Compute as needed: little memory, slow (many cache misses)
- ☑ Block coordinate descent: **fast as possible given available memory (few cache misses)**

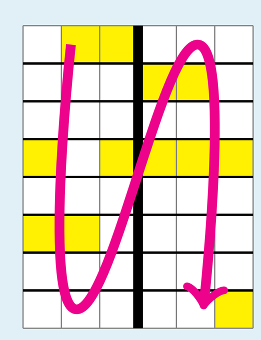
Block-wise strategy for Λ



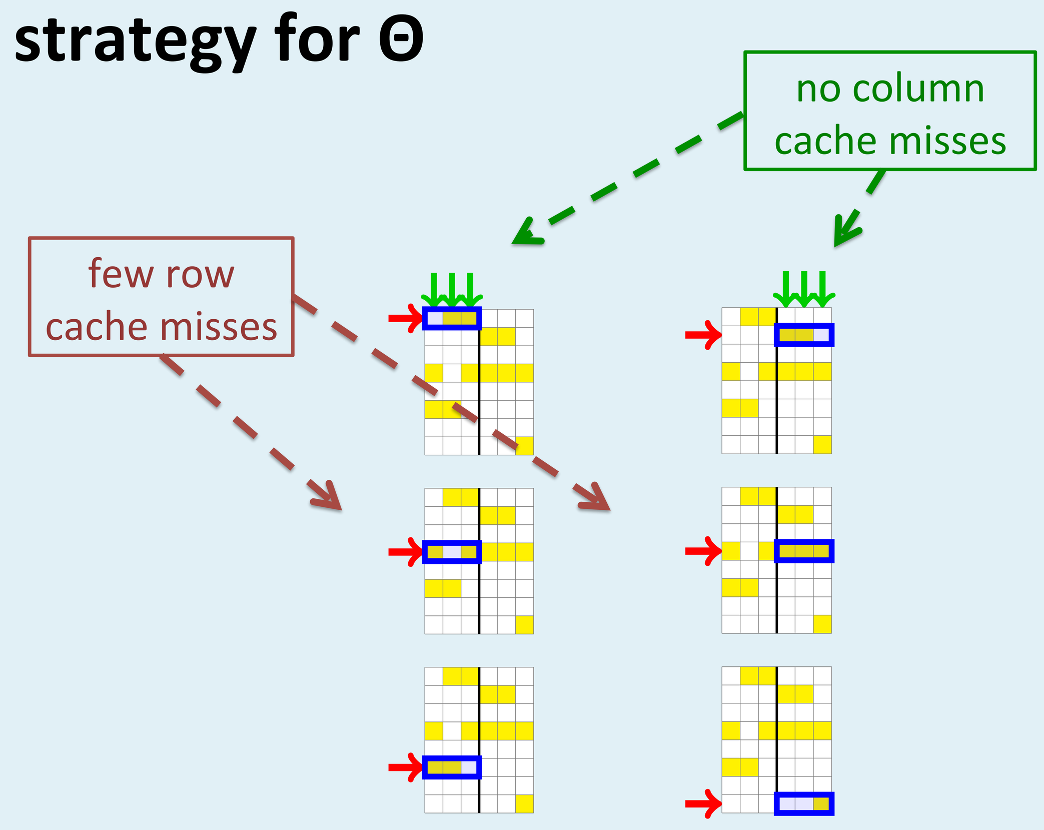
- Share computations across groups of rows
- Choose blocks via graph clustering on **Λ**



Block-wise strategy for Θ

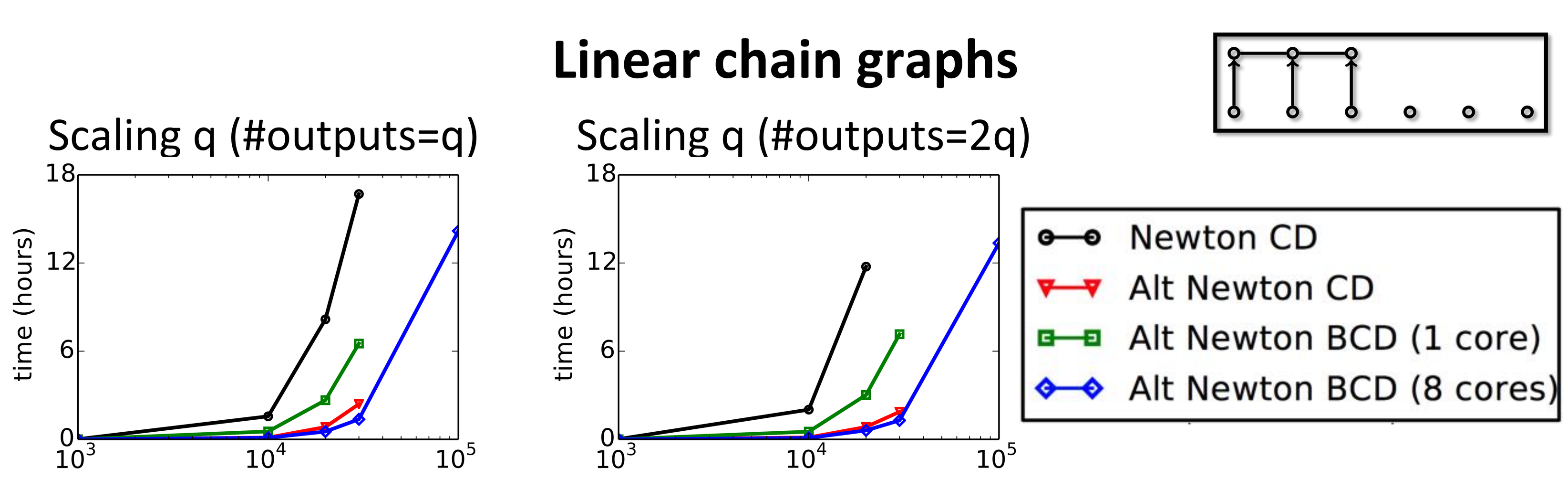


- Share computations across groups of columns
- Choose blocks via graph clustering on **$\Theta^T\Theta$**
- Exploit row-wise sparsity

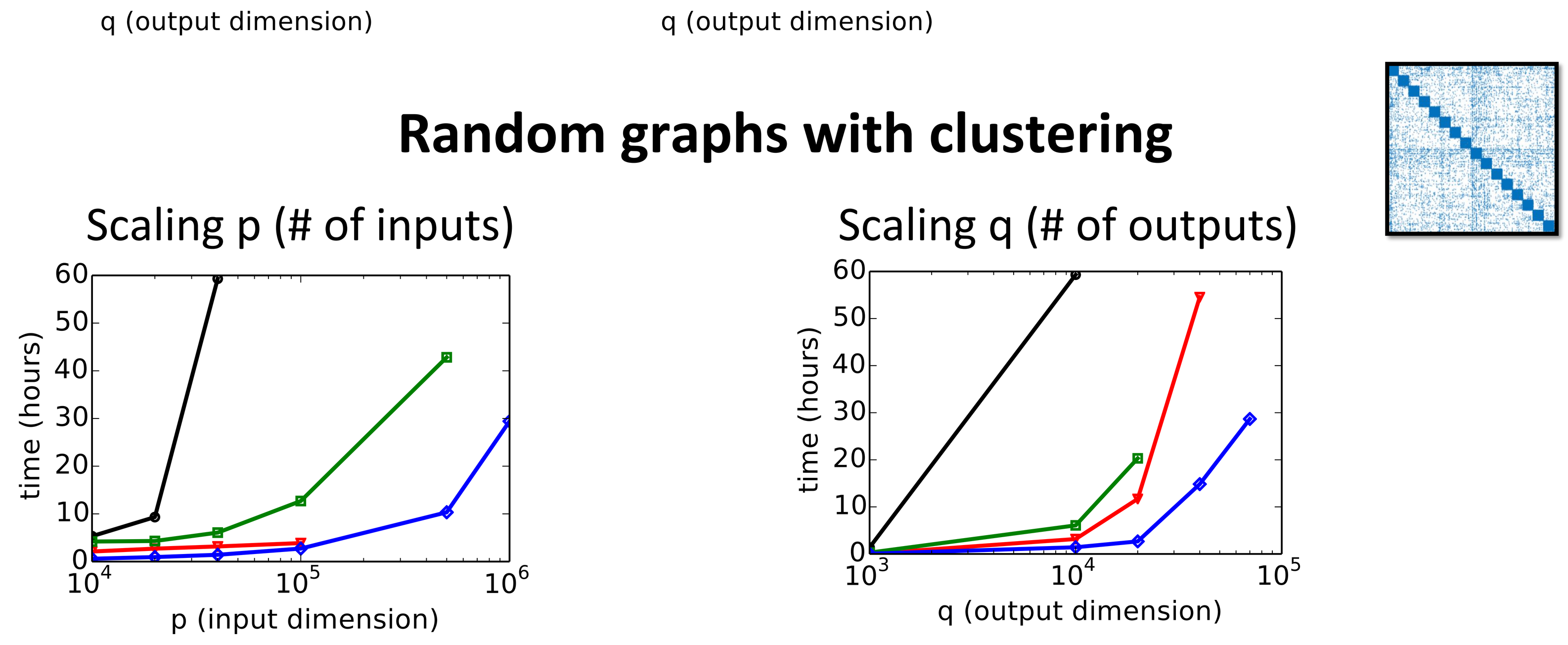


Simulation Results

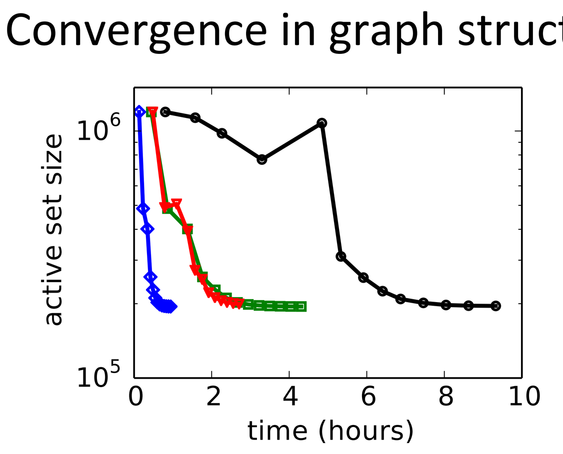
Linear chain graphs



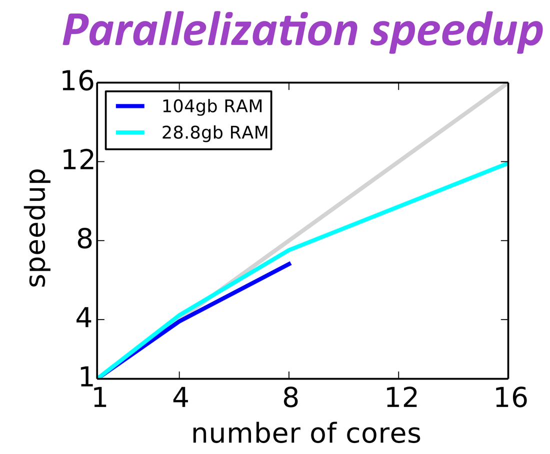
Random graphs with clustering



Convergence in graph structure



Parallelization speedup

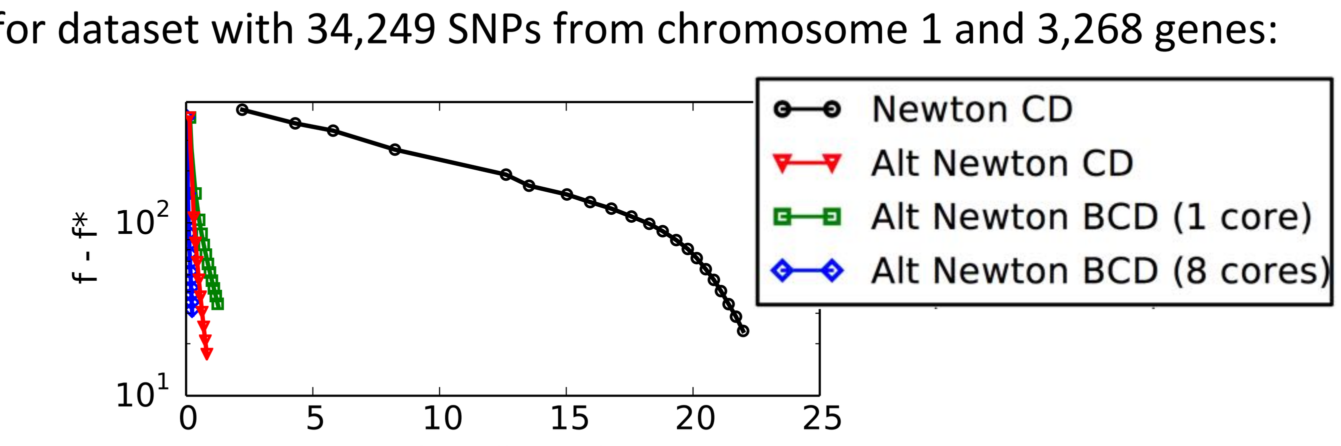


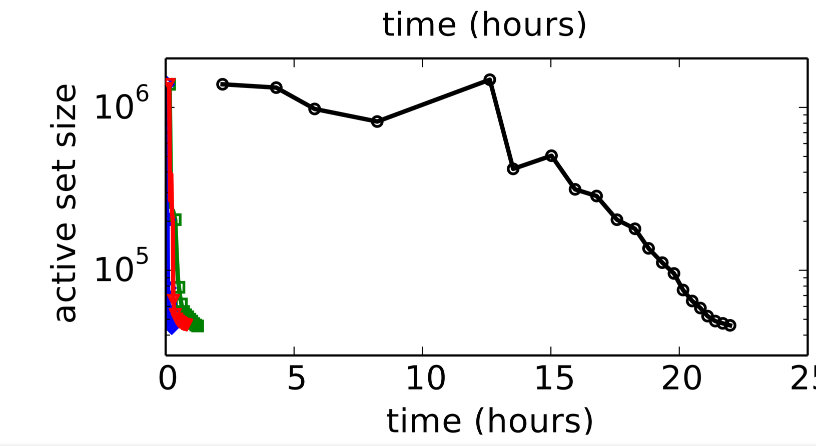
Genomic Data Analysis

- Genotypes and gene expression levels for 171 individuals with asthma
- Contains 442,440 SNPs and 10,256 genes with variance > 0.01
- Regularization parameters chosen to learn graphs with 10q edges

p	q	Newton CD	Alt Newton CD	Alt Newton BCD
34,249	3,268	22.0	0.51	**0.24**
34,249	10,256	> 50	2.4	**2.3**
442,440	3,268	*	*	**11**

Results for dataset with 34,249 SNPs from chromosome 1 and 3,268 genes:





References

[1] Sohn & Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. AISTATS 2012.
[2] Yuan & Zhang. Partial gaussian graphical model estimation. IEEE Transactions on Information Theory. 2014.
[3] Wytock & Kolter. Sparse Gaussian conditional random fields: algorithms, theory, and application to energy forecasting. ICML 2013.
[4] Hsieh et al. BIG & QUIC: Sparse inverse covariance estimation for a million variables. NIPS 2013.