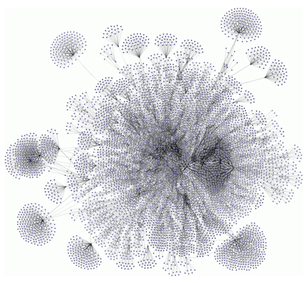


High Dimensional Sparse Inverse Covariance Estimation using Greedy Methods

Christopher C. Johnson, Ali Jalali, and Pradeep Ravikumar

cjohnson@cs.utexas.edu, alij@mail.utexas.edu, pradeepr@cs.utexas.edu

THE UNIVERSITY OF TEXAS AT AUSTIN

Inverse Covariance Estimation

- Random Variables $X_1, \ldots, X_p \sim \mathcal{N}(0, \Sigma)$
- PDF $f(x_1, \ldots, x_n; \theta^*) = \frac{\exp\left\{-\frac{1}{2} x^T \Theta^* x\right\}}{\sqrt{(2\pi)^p \det(\Theta^*)^{-1}}}$
- Sparse Inverse Covariance Matrix $\Theta^* = (\Sigma^*)^{-1}$ (number of off-diagonal non-zeros is small)
- iid Observations $x_1^n = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$
- Goal: Estimate sparse Θ given $n \ll p$ iid samples.
- Sample Covariance Matrix

$$\widehat{\Sigma}^n := \frac{1}{n} \sum_{k=1}^{n} x^{(k)} (x^{(k)})^T$$

- Gaussian Log-Likelihood Loss

$$\mathcal{L}(\Theta; \widehat{\Sigma}^n) := \langle\langle \Theta, \ \widehat{\Sigma}^n \rangle\rangle - \log\det(\Theta)$$

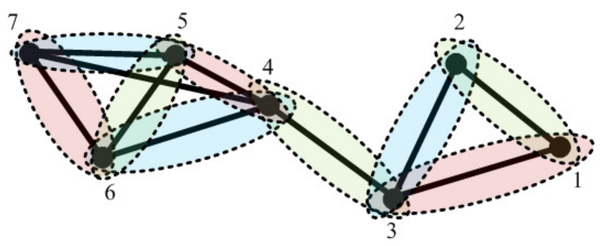
- Would like to solve sparsity constrained MLE

$$\widehat{\Theta} = \arg \min_{\|\Theta\|_0 \leq s} \mathcal{L}(\Theta; \widehat{\Sigma}^n).$$

Cannot solve directly as optimization is non convex!

Gaussian Graphical Models

- Every Gaussian distribution can be represented as a pairwise Gaussian Markov Random Field (GMRF)
- Sparse Graph $\mathcal{G}^*(\mathcal{V}^*, \mathcal{E}^*)$ where vertices correspond to variables and edges correspond to off-diagonal, non-zero inverse covariates $E(\Theta^*) := \{i, j \in V \mid i \neq j, \Theta^*_{ij} \neq 0\}$



- Estimating non-zero elements of $\Theta^* = (\Sigma^*)^{-1}$ equivalent to estimating sparse set of edges $E(\Theta^*) := \{i, j \in V \mid i \neq j, \Theta^*_{ij} \neq 0\}$
- GLasso [3] solves ℓ_1-regularized Gaussian MLE

$$\widehat{\Theta} := \arg \min_{\Theta \in \mathcal{S}^p_{++}} \left\{ \langle\langle \Theta, \ \widehat{\Sigma}^n \rangle\rangle - \log\det(\Theta) + \lambda_n \|\Theta\|_{1,\text{off}} \right\}$$

- Neighborhood Lasso [2] estimates neighborhood of each variable X_r independently using ℓ_1-regularized least squares loss

$$\Gamma_r = \arg\min_{\Gamma_r} \frac{1}{2n} \sum_{i=1}^{n} \left(x_r^{(i)} - \sum_{t \neq r} \Gamma_{rt} x_t^{(i)} \right)^2 + \lambda_n \, ||\Gamma_r||_1$$

Sparsistency Guarantees

Theorem 1 (Global Greedy Sparsistency). *Assuming $\Theta^*_{min} = \Omega(1/\sqrt{d})$ and $n = O(d \log(p))$ then running Algorithm 1 (global greedy) with stopping threshold $\epsilon_S = \Omega(d \log(p)/n)$ ensures that the full structure of the model is recovered ($\widehat{E} = E^*$) with high probability.*

Theorem 2 (Neighborhood Greedy Sparsistency). *Assuming $\min_{t \in \mathcal{N}^*(r)} \Gamma^*_{rt} = \Omega(1/\sqrt{d})$ and $n = O(d \log(p))$ then running Algorithm 2 (neighborhood greedy) with stopping threshold $\epsilon_S = \Omega(d \log(p)/n)$ ensures that the full structure of the model is recovered ($\widehat{E} = E^*$) with high probability.*

References

[1] A. Jalali, C. Johnson, P. Ravikumar On Learning Discrete Graphical Models Using Greedy Methods In NIPS, 2011

[2] N. Meinshuasen, P. Bühlmann High-dimensional graphs and variable selection with the Lasso In Annals of Statist., 2006

[3] P. Ravikumar, M. Wainwright, G. Raskutti, B. Yu High-dimensional Covariance Estimation by Minimizing ℓ_1-penalized log-determinant Divergence In Electron J. Statist., 2010.

Global Greedy Method

Algorithm 1 (Global greedy method).

Let $\mathcal{L}(\Theta) = \langle\langle \Theta, \ \widehat{\Sigma}^n \rangle\rangle - \log\det(\Theta)$

Initialize $\widehat{\Theta}^{(0)} \leftarrow \mathbb{I}$ *and* $k \leftarrow 1$
while *true* ***do*** {**Forward Step**}
  Choose best new edge $\widehat{e}_{i_* j_*}$ *according to* $\mathcal{L}(\widehat{\Theta}^{(k-1)})$;
  if *Decrease in loss* (δ_f^k) *of adding* $\widehat{e}_{i_* j_*}$ *to* $\widehat{\Theta}^{(k-1)}$ *is* $\leq \epsilon_S$ ***then***
    break;
  end if
  Add $\widehat{e}_{i_* j_*}$ *to* $\widehat{\Theta}^{(k-1)}$;
  Optimize edge weights;
  $k \leftarrow k + 1$;

  while *true* ***do*** {**Backward Step**}
    Choose weakest edge $\widehat{e}_{i^* j^*}$ *in* $\widehat{\Theta}^{(k-1)}$ *according to* $\mathcal{L}(\widehat{\Theta}^{(k-1)})$;
    if *Increase in loss of removing* $\widehat{e}_{i^* j^*}$ *from* $\widehat{\Theta}^{(k-1)}$ *is* $> \nu \delta_f^{(k)}$ ***then***
      break;
    end if
    Remove $\widehat{e}_{i^* j^*}$ *from* $\widehat{\Theta}^{(k-1)}$;
    Optimize edge weights;
    $k \leftarrow k - 1$;
  end while
end while

Neighborhood Greedy Method

Algorithm 2 (Neighborhood greedy method).

Let $\mathcal{L}(\Gamma) = \frac{1}{2n} \sum_{i=1}^{n} \left(x_r^{(i)} - \sum_{t \neq r} \Gamma_{rt} x_t^{(i)} \right)^2$

Initialize $\widehat{\Gamma}_r^{(0)} \leftarrow \mathbf{0}$ *and* $k \leftarrow 1$
while *true* ***do*** {**Forward Step**}
  Choose best new edge $\widehat{e}_{r t_*}$ *according to* $\mathcal{L}(\widehat{\Gamma}_r^{(k-1)})$;
  if *Decrease in loss* (δ_f^k) *of adding* $\widehat{e}_{r t_*}$ *to* $\widehat{\Gamma}_r^{(k-1)}$ *is* $\leq \epsilon_S$ ***then***
    break;
  end if
  Add $\widehat{e}_{r t_*}$ *to* $\widehat{\Gamma}_r^{(k-1)}$;
  Optimize edge weights;
  $k \leftarrow k + 1$;

  while *true* ***do*** {**Backward Step**}
    Choose weakest edge $\widehat{e}_{r t^*} \in \widehat{\Gamma}_r^{(k-1)}$ *according to* $\mathcal{L}(\widehat{\Gamma}_r^{(k-1)})$;
    if *Increase in loss of removing* $\widehat{e}_{r t^*}$ *from* $\widehat{\Gamma}_r^{(k-1)}$ *is* $> \nu \delta_f^{(k)}$ ***then***
      break;
    end if
    Remove $\widehat{e}_{r t^*}$ *from* $\widehat{\Gamma}_r^{(k-1)}$;
    Optimize edge weights;
    $k \leftarrow k - 1$;
  end while
end while

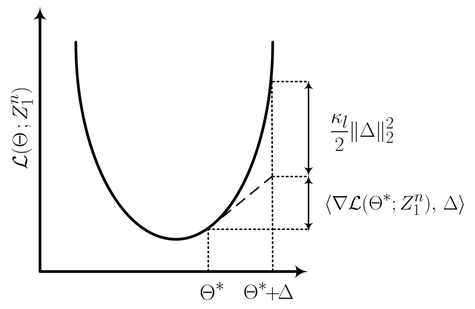
Model Assumptions: Restricted Strong Convextiy/Smoothness

- (A1) Restricted Strong Convexity (RSC)
  - Restricted strong convexity RSC(s) with parameter κ_l holds iff for any s-sparse vector Δ, we have

$$\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) \geq \langle \nabla \mathcal{L}(\Theta^*), \ \Delta \rangle + \frac{\kappa_l}{2} \|\Delta\|_2^2$$

- (A2) Restricted Strong Smoothness (RSS)
  - Restricted strong smoothness RSS(s) with parameter κ_u holds iff for any s-sparse vector Δ, we have

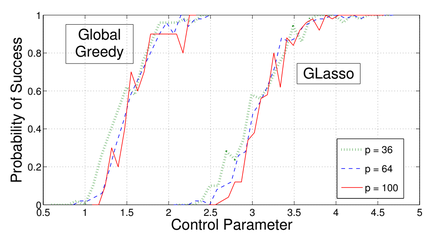
$$\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) \leq \langle \nabla \mathcal{L}(\Theta^*), \ \Delta \rangle + \frac{\kappa_u}{2} \|\Delta\|_2^2$$



Experimental Results

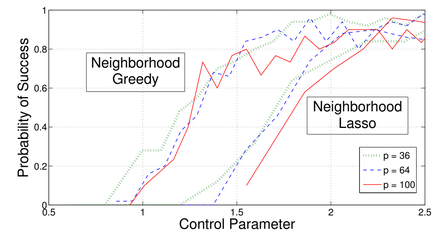
True Graphical Model

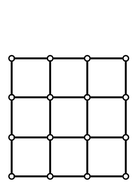
Probability of Success

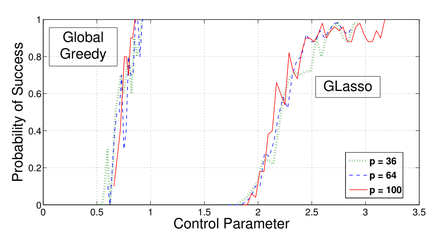


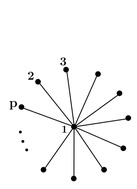
True Graphical Model

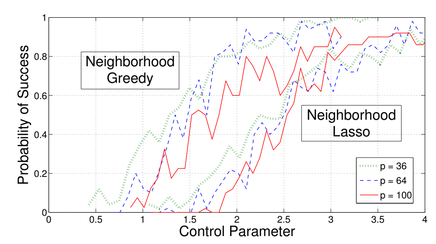
Probability of Success











Comparison of GMRF Structure Learning Methods

	Model Assumptions	Sample Complexity	Min Non-Zero Values
Graphical Lasso	Irrepresentability	$O(d^2 \log(p))$	$\Omega(1/d)$
Neighborhood Lasso	Irrepresentability	$O(d \log(p))$	$\Omega(1/\sqrt{d})$
Global Greedy	RSC/RSS	$O(d \log(p))$	$\Omega(1/\sqrt{d})$
Neighborhood Greedy	RSC/RSS	$O(d \log(p))$	$\Omega(1/\sqrt{d})$

State of the art in all areas!