

# Partial Correlation Estimation by Joint Sparse Regression Models

Ji Zhu, U Michigan

Joint work with Jie Peng (UC Davis), Pei Wang (Fred Hutchinson) and Nengfeng Zhou (U Michigan)

# Covariance Selection

- Identification and estimation of non-zero entries in the **concentration matrix** (Dempster, 1972).
- Useful in elucidating associations among a set of random variables.

# Outline

- Background
- Joint sparse regression
- Simulation results
- Real application
- Asymptotic results
- Summary

# Problem Setup

- Data:  $\{X_j^i : j = 1, \dots, p; i = 1, \dots, n\}$ 
  - The  $i$ th sample:  $\mathbf{X}^i = (X_1^i, \dots, X_p^i)$
  - We assume

$$\mathbf{X}^1, \dots, \mathbf{X}^n \text{ i.i.d. } \sim \mathbf{N}(\mathbf{0}, \Sigma)$$

- Let  $\Sigma^{-1} = (\sigma^{jj'})$  (concentration matrix), then

$$\rho^{jj'} = \text{Cor}(X_j, X_{j'} | X_{-\{j, j'\}}) = -\frac{\sigma^{jj'}}{\sqrt{\sigma^{jj} \sigma^{j'j'}}}.$$

- Goal: estimate the partial correlation  $\rho^{jj'}$ , especially identify the non-zero ones.

# Motivating Example: Genetic Regulatory Networks

- Gene-gene interactions play an important role in biological process.
- Strong interactions  $\implies$  significant correlations among mRNA expression levels.
- High throughput microarray technique enables us to monitor mRNA expression levels for tens of thousands of genes simultaneously in one experiment.
- The correlation structure inferred from expression arrays helps to cast light on the genetic regulatory network.

# Gaussian Graphical Model

- Specifically, assume the expression levels of  $p$  genes

$$\mathbf{X} = (X_1, \dots, X_p)^\top \sim \mathbf{N}(\mathbf{0}, \Sigma).$$

- Define an **undirected graph** according to  $\Sigma$ 
  - Each node represents a gene.
  - There is an edge connecting node  $j$  and node  $j'$  **if and only if**

$$\text{Cor}(X_j, X_{j'} | X_{-\{j, j'\}}) \neq 0.$$

# High Dimensionality

- Challenge:  $p \gg n$ . For example, for gene expression data:
  - Often  $n \sim 10^2$  and  $p \sim 10^3$ .
  - Need to estimate about  $p(p - 1)/2 \approx 10^6$  parameters with only  $10^2$  samples.

# Important Properties of Networks

- Sparsity
  - Many real-life complex networks (including GRNs) are intrinsically sparse.
  - There are  $p(p - 1)/2$  parameters, but the majority of them are zero.
- Hubs
  - Nodes that are connected to many other nodes.
  - The existence of hubs is a well known phenomenon for many large networks.

## Existing Methods

- Dobra et al. (2004): Bayesian approach based on Cholesky decomposition.
- Li and Gui (2007): threshold gradient descent regularization procedure.

# Penalized Likelihood

Yuan and Lin (2007) propose a [penalized maximum likelihood](#) approach by minimizing

$$-\log \det (\boldsymbol{\Sigma}^{-1}) + \text{trace}(\boldsymbol{\Sigma}^{-1}S) + \lambda \sum_{j \neq j'} |\sigma^{jj'}|,$$

where  $S$  denotes the sample covariance matrix.

- $L_1$ -norm: [LASSO](#) penalty.
- Use an interior point algorithm (very slow).
- Friedman et al. (2007) develop a very efficient algorithm — [glasso](#).

# Least Squares Regression

- Predictor variables (covariates, inputs):

$$\mathbf{X} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$$

- Response variable (outcome, output):

$$Y \in \mathbb{R}$$

- Ordinary least squares estimates (OLS)

$$\min_{\beta_j} \sum_{i=1}^n \left( Y_i - \sum_{j=1}^p \beta_j X_j^i \right)^2$$

Two reasons why not satisfied with the least squares estimate:

- Prediction accuracy

$$\begin{aligned} \mathbb{E}\left((\hat{Y} - f)^2\right) &= \left(\mathbb{E}(\hat{Y}) - f\right)^2 + \text{Var}(\hat{Y}) \\ \text{MSE} &= \text{Bias}^2 + \text{Variance} \end{aligned}$$

### Bias-variance trade-off

- Parsimony: we would like to determine a smaller subset of predictors that exhibit the strongest effects.

# LASSO

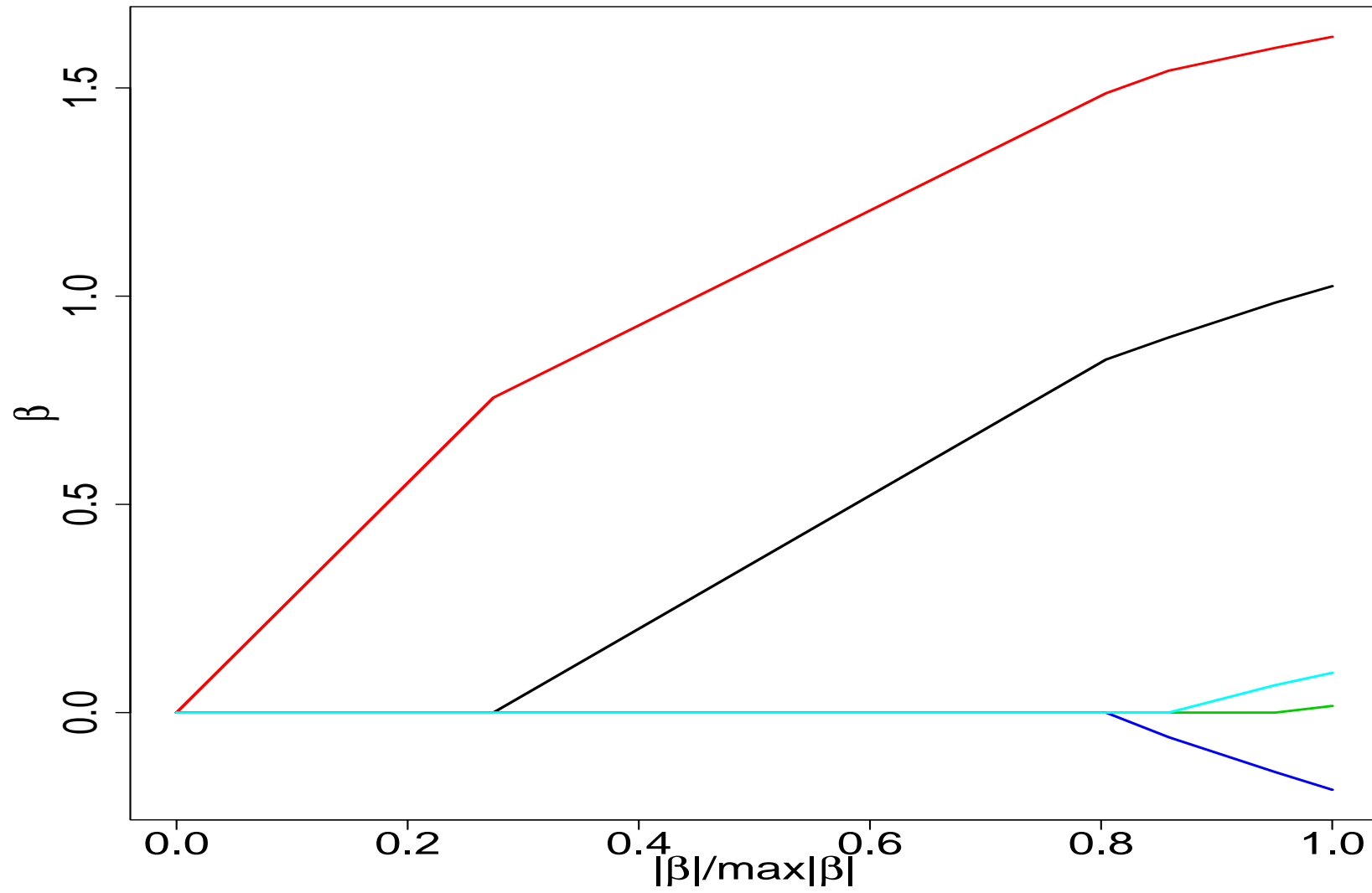
Least absolute shrinkage and selection operator (Chen, Donoho and Saunders 1996; Tibshirani 1996)

$$\min_{\beta_j} \sum_{i=1}^n \left( Y_i - \sum_{j=1}^p \beta_j X_j^i \right)^2 + \lambda (|\beta_1| + |\beta_2| + \cdots + |\beta_p|)$$

- Shrinkage
- **Sparsity**: some fitted coefficients are **exactly** zero

Continuous variable selection

# LASSO Path



# Outline

- ✓ Background
  - Joint sparse regression
    - Model
    - Algorithm
  - Simulation results
  - Real application
  - Asymptotic results
  - Summary

## Connection with Regression

- $X_j$  is expressed as

$$X_j = \sum_{j' \neq j} \beta_{jj'} X_{j'} + \epsilon_j$$

such that  $\epsilon_j$  is independent of  $X_{-j}$  if and only if  $\beta_{jj'} = -\sigma^{jj'} / \sigma^{jj}$ . Furthermore,

$$\beta_{jj'} = \rho^{jj'} \sqrt{\frac{\sigma^{j'j'}}{\sigma^{jj}}}.$$

- We also have  $\text{Var}(\epsilon_j) = 1/\sigma^{jj}$ .

# Joint Sparse Regression Model

- Propose to minimize the joint criterion

$$\sum_{j=1}^p \left\| \mathbf{X}_j - \sum_{j' \neq j} \rho^{jj'} \sqrt{\sigma^{j'j'} / \sigma^{jj}} \mathbf{X}_{j'} \right\|^2 + \lambda \sum_{j \neq j'} |\rho^{jj'}|$$

where  $\mathbf{X}_j = (X_j^1, \dots, X_j^n)^\top$ .

- **Space**: sparse partial correlation estimation.

# Neighborhood Selection

Meinshausen and Bühlmann (2006)

- Fit  $p$  **individual** lasso regressions

$$\min_{\beta_{jj'}} \left\| \mathbf{X}_j - \sum_{j' \neq j} \beta_{jj'} \mathbf{X}_{j'} \right\|^2 + \lambda \sum_{j' \neq j} |\beta_{jj'}|, \quad j = 1, \dots, p$$

- Calculate  $\rho^{jj'} = \text{sign}(\beta_{jj'}) \sqrt{\beta_{jj'} \beta_{j'j}}$ .

## Space vs MB

- **space** assures the sign consistency between  $\beta_{jj'}$  and  $\beta_{j'j}$ , as it **estimates**  $\{\rho^{jj'}\}$  directly. While neighborhood selection can lead to **contradictory neighborhoods**. This also reduces the number of unknown parameters almost by half.
- In **space**, sparsity is utilized for the partial correlations as a **whole view**. However, in the neighborhood selection approach, sparsity is imposed on **each neighborhood**. The former treatment is more effective for networks with **hubs**.

# Weights

$$\sum_{j=1}^p w_j \left\| \mathbf{X}_j - \sum_{j' \neq j} \rho^{jj'} \sqrt{\sigma^{j'j'} / \sigma^{jj}} \mathbf{X}_{j'} \right\|^2 + \lambda \sum_{j \neq j'} |\rho^{jj'}|.$$

- Equal weights:  $w_j = 1$  (space)
- Residue variance based weights:  $w_j = \sigma^{jj}$  (space.sw)
- Degree based weights:  $w_j$  is proportional to the estimated degree for  $X_j$  (space.dew)

# Outline

- ✓ Background
  - Joint sparse regression
    - ✓ Model
      - Algorithm
  - Simulation results
  - Real application
  - Asymptotic results
  - Summary

# Algorithm

Iterative approach

$$\min_{\rho^{jj'}, \sigma^{jj}} \sum_{j=1}^p w_j \left\| \mathbf{X}_j - \sum_{j' \neq j} \rho^{jj'} \sqrt{\sigma^{j'j'} / \sigma^{jj}} \mathbf{X}_{j'} \right\|^2 + \lambda \sum_{j \neq j'} |\rho^{jj'}|$$

- Given  $\sigma_{(k)}^{jj}$ , estimate  $\rho_{(k)}^{jj'}$  — **LASSO regression**
- Given  $\sigma_{(k)}^{jj}$  and  $\rho_{(k)}^{jj'}$ , estimate  $\sigma_{(k+1)}^{jj}$ :

$$\sigma_{(k+1)}^{jj} = 1 / \widehat{\text{Var}}(\hat{\epsilon}_j),$$

where  $\hat{\epsilon}_j = \mathbf{X}_j - \sum_{j' \neq j} \rho_{(k)}^{jj'} \sqrt{\sigma_{(k)}^{j'j'} / \sigma_{(k)}^{jj}} \mathbf{X}_{j'}$ .

# LASSO Regression

$$\min \|\mathcal{Y} - \mathcal{X}\boldsymbol{\rho}\|^2 + \lambda\|\boldsymbol{\rho}\|_1$$

- $\mathcal{Y}$  is  $np \times 1$ , and  $\mathcal{X}$  is  $np \times p(p-1)/2$  (sparse).
- We develop a new algorithm, [active-shooting](#), for solving the lasso-type problem.

## Idea of Active-Shooting

- Shooting (Fu, 1998; Friedman et al., 2007): update each coordinate iteratively; computationally competitive compared with LARS/LASSO (Efron et al. 2004).

$$\begin{array}{cccc} & & \vdots & \\ & & & \\ & & & \\ \beta_1 & \hat{\beta}_2 & \cdots & \hat{\beta}_p \\ \hat{\beta}_1 & \beta_2 & \cdots & \hat{\beta}_p \\ & & \vdots & \\ \hat{\beta}_1 & \hat{\beta}_2 & \cdots & \beta_p \\ & & \vdots & \end{array}$$



## Active-Shooting vs Shooting

- Apply both algorithms to a simulation example in Friedman et al. (2007).
- $n = 100$ .
- $p_0$ : number of non-zero coefficients.
- Record the **number of iterations**.

$p$	$p_0$	shooting	active-shooting
200	14	29600	4216
500	25	154000	10570
1000	28	291000	17029

## Computational Cost of Space

- Computational cost of `space` is  $O(np^2)$ .
- Takes about 30 seconds to fit a data set with  $p = 1000$  and  $n = 200$  (1000 true edges) on a linux server with AMD Opteron 2.6GHz CPU and 4G RAM.
- R package `space` is available on cran.

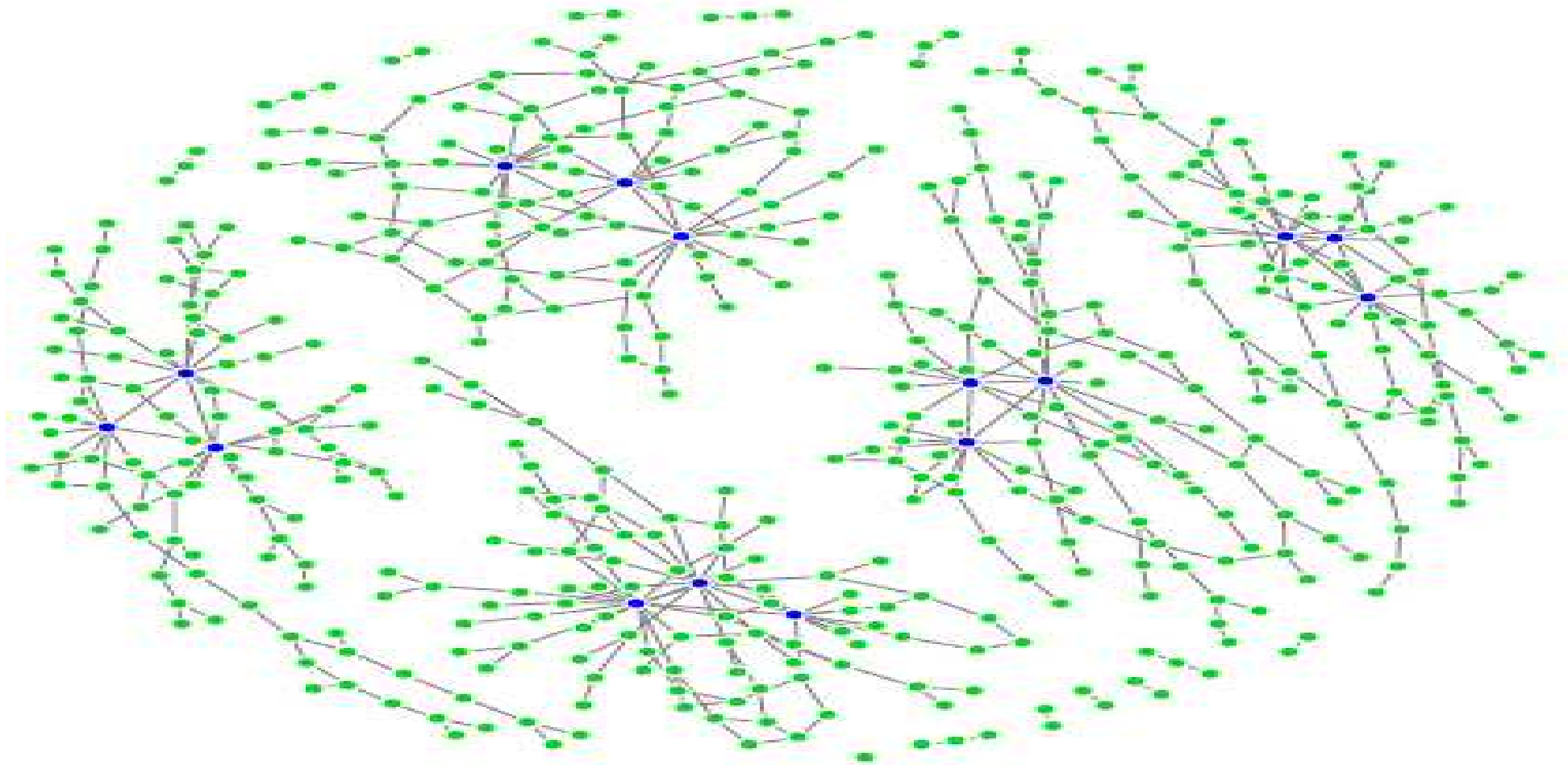
# Outline

- ✓ Background
- ✓ Joint sparse regression
  - Simulation results
  - Real application
  - Asymptotic results
  - Summary

## Simulation Setting

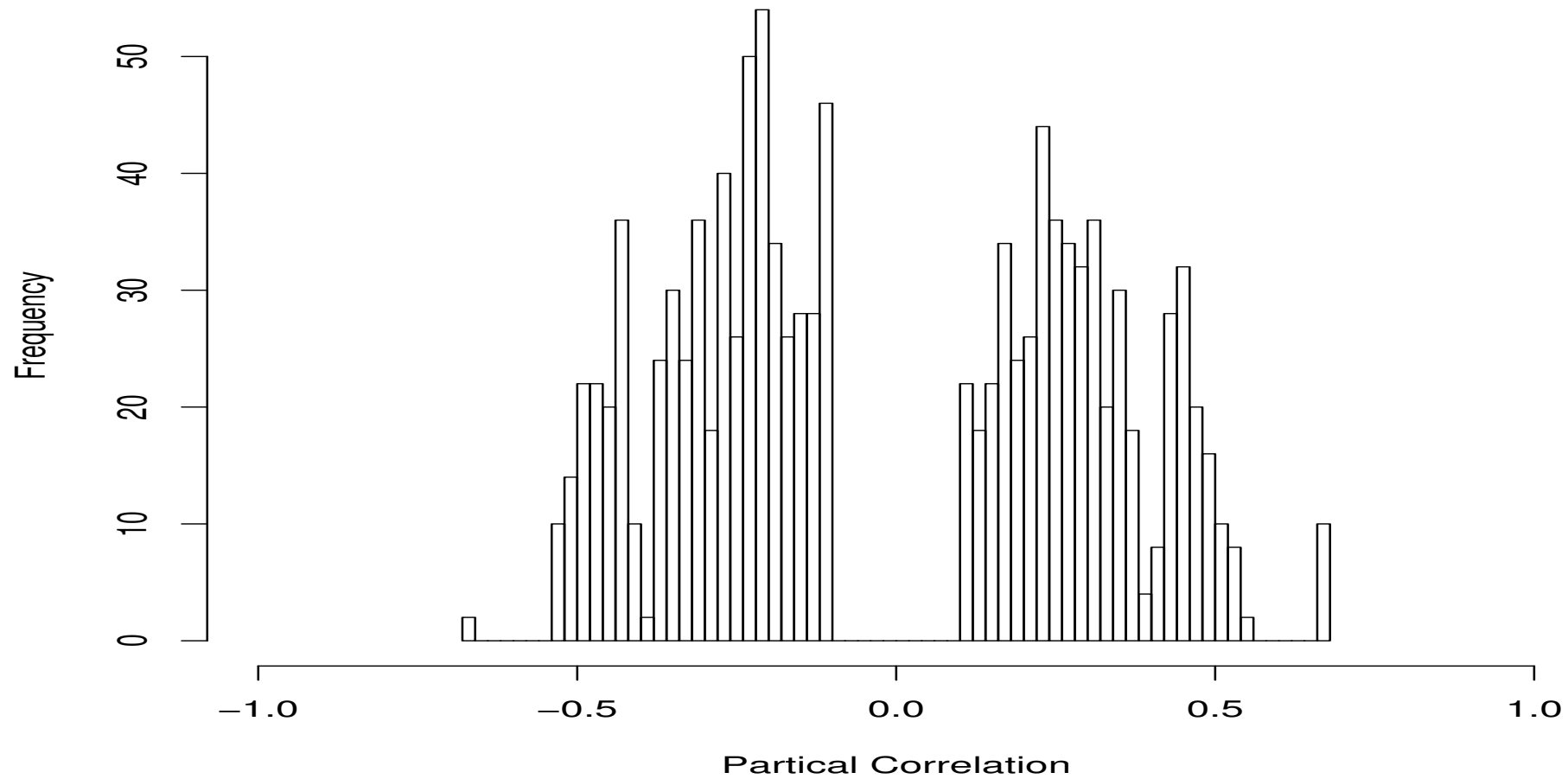
- Simulate a network of size  $p$  from a given degree distribution.
- Simulate a concentration matrix  $\Sigma^{-1}$  according to the network topology.
- Simulate  $\mathbf{X}^1, \dots, \mathbf{X}^n$  i.i.d.  $\sim \mathbf{N}(\mathbf{0}, \Sigma)$ .
- Compare the performances of three methods: `space`, `MB`, and `glasso`.
- Evaluate the methods at a series of values of  $\lambda$ .

# Example: Hub Network

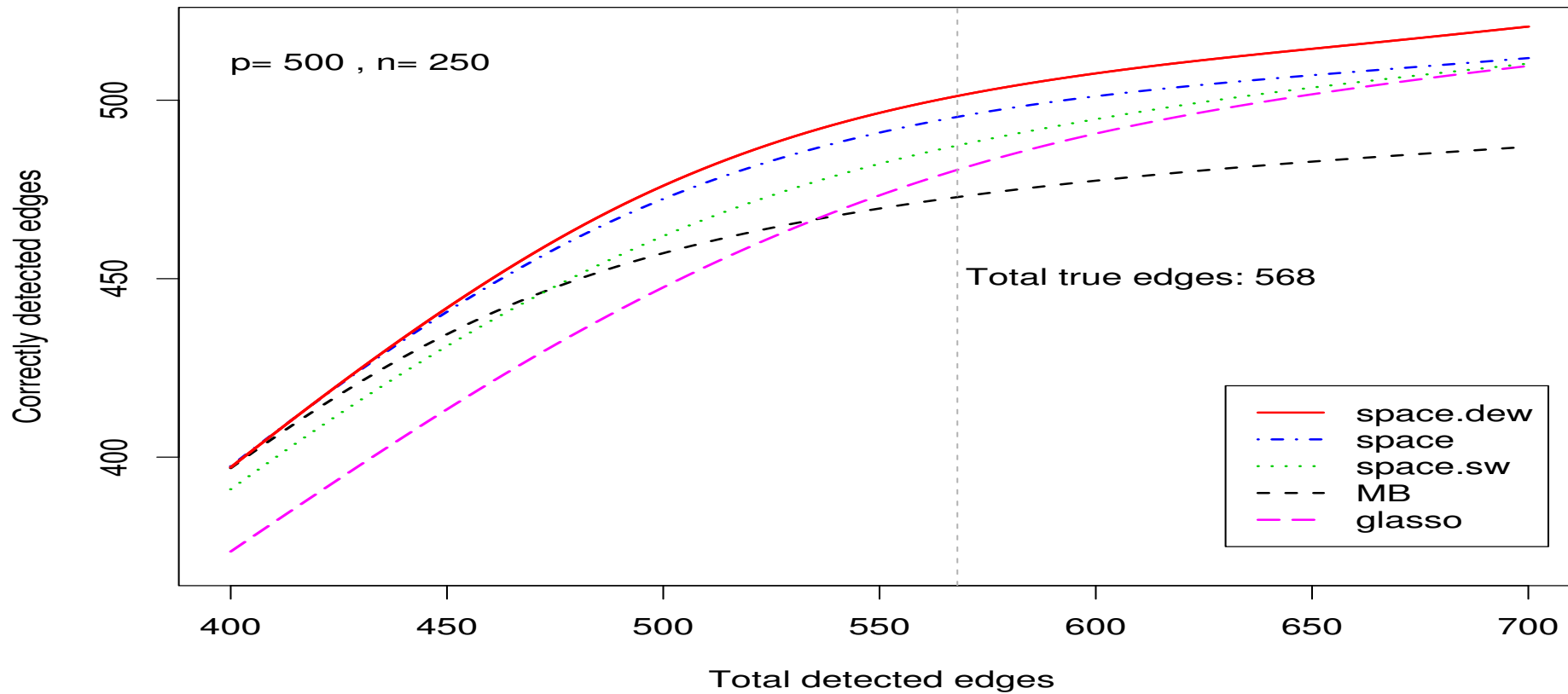


$p = 500$  nodes and 568 edges; 15 hub nodes (in blue)

# Example: Partial Correlation

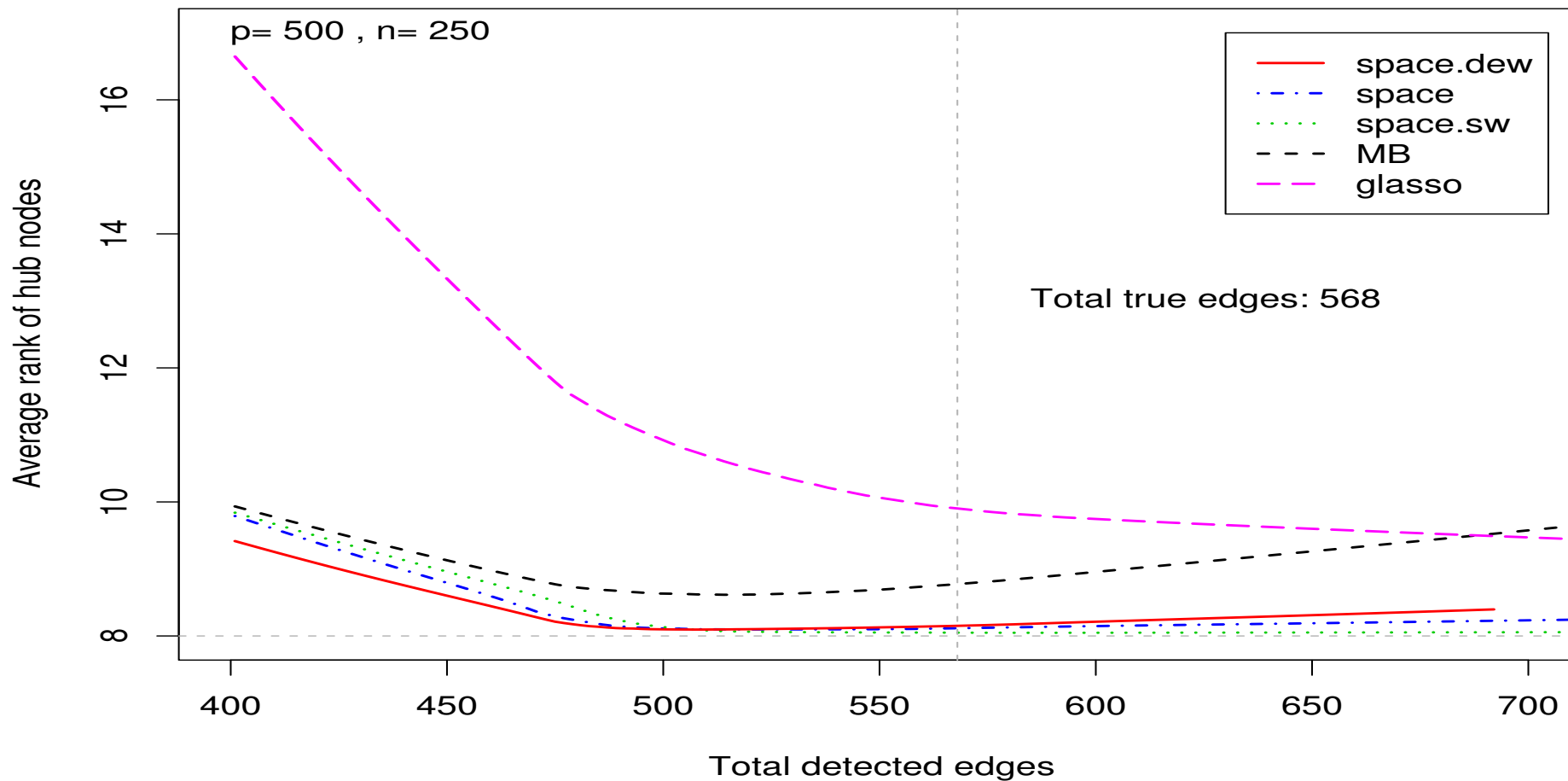


# True Positive



The number of correctly detected edges vs the number of total detected edges (averaged over 50 repetitions).

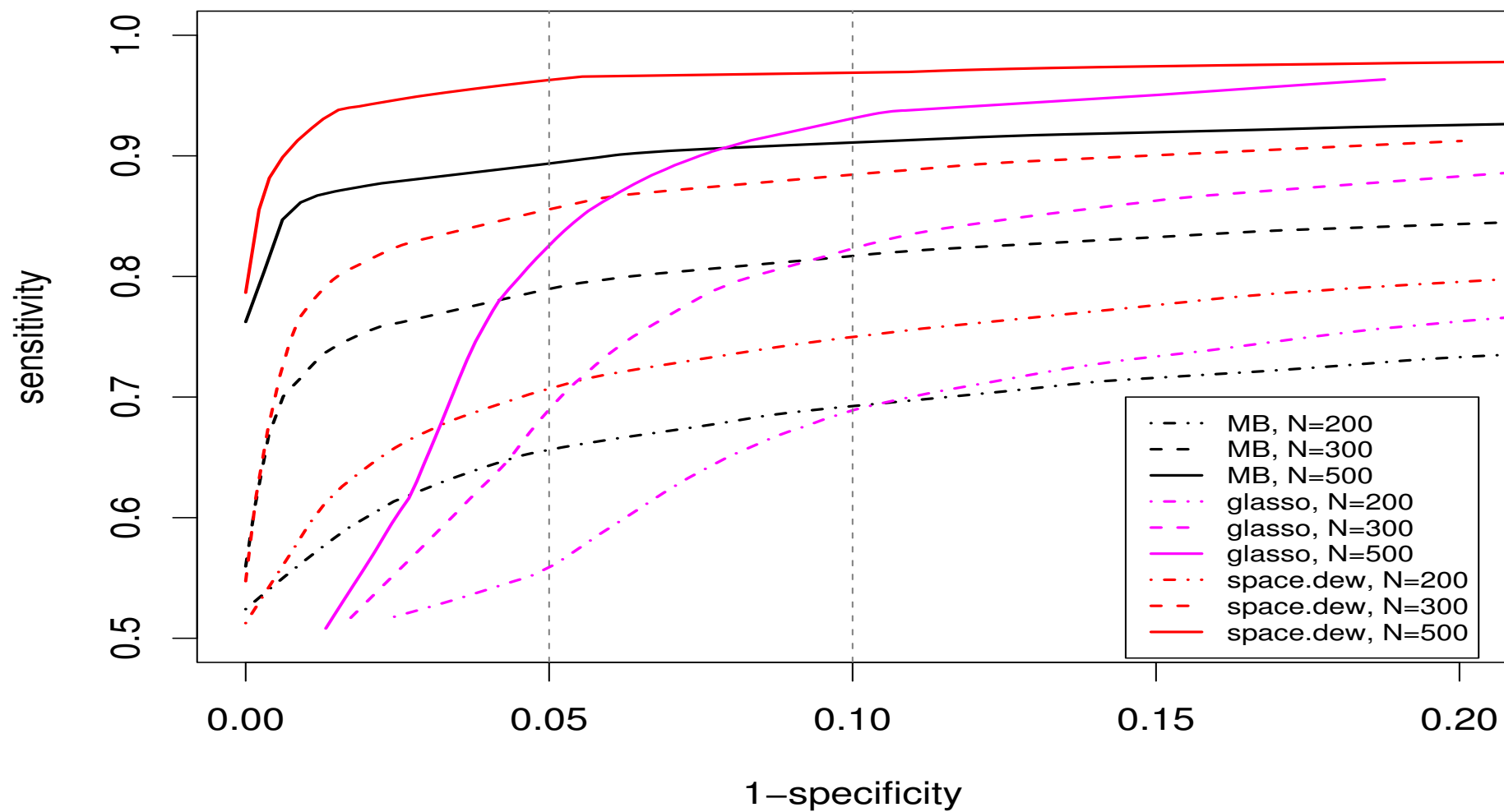
# Hub Nodes



Average rank of the 15 true hub nodes.

## Different Sample Size

- $p = 1000$  with 1163 true edges
- $n = 200, 300, 500$



## Sensitivity

Power in identifying correct edges when FDR is controlled at 0.05.

$p$	$n$	MB	glasso	space.dew
500	250	0.784	0.655	0.844
	200	0.656	0.559	0.707
1000	300	0.790	0.690	0.856
	500	0.894	0.826	0.963

## Remarks

- Other networks: power-law network with the power law parameter  $\alpha = 2.3$ ; uniform network as in Meinshausen and Buhlmann (2006); AR(2) network as in Friedman et al. (2007).
- `space` performs favorably over both MB and `glasso` in both [edge detection](#) and [hub identification](#).
- The advantage of `space` is more obvious in networks with hubs (e.g., the hub-network and the power-law network) than networks without hubs.
- Sample size is important.

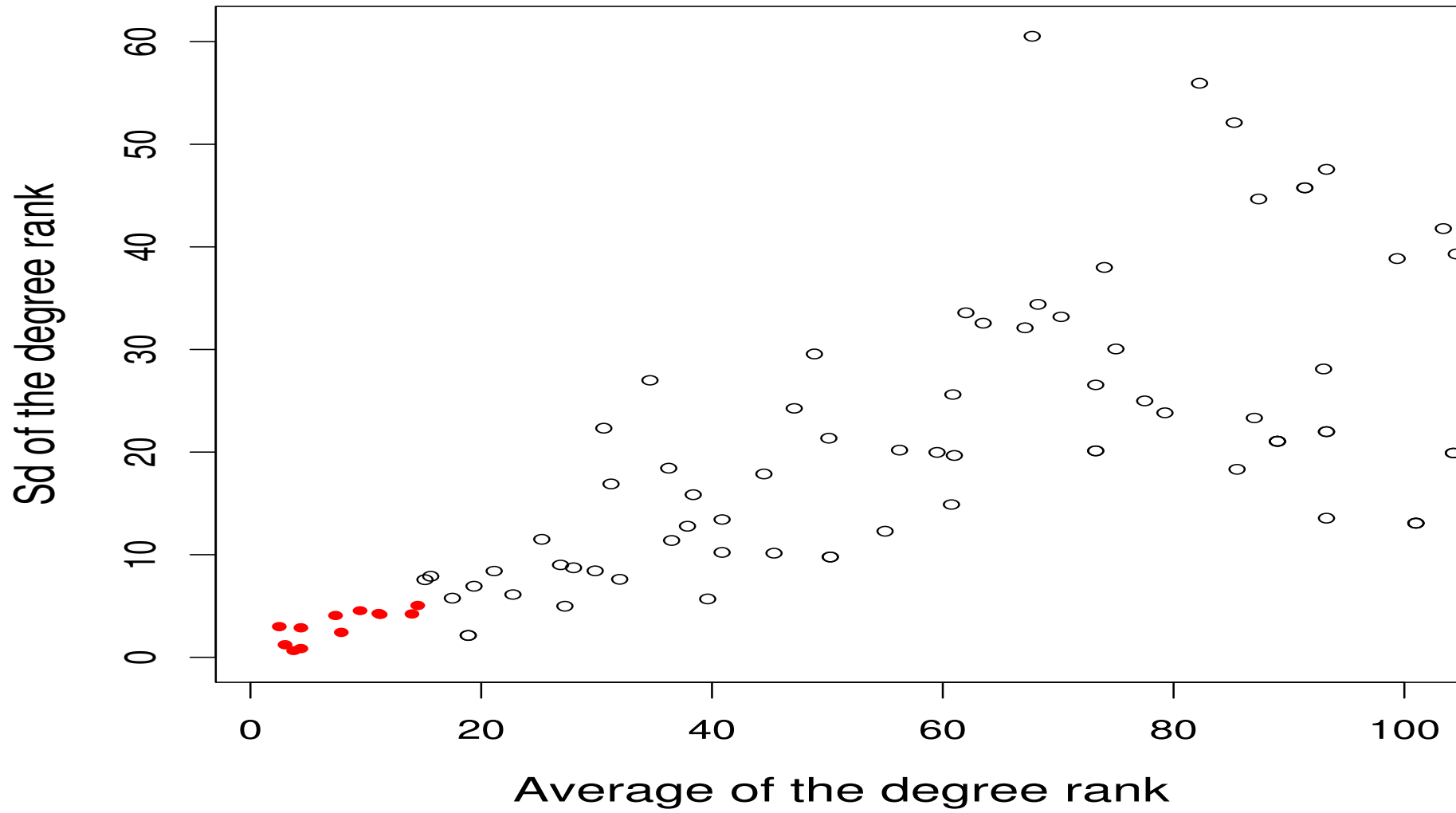
# Outline

- ✓ Background
- ✓ Joint sparse regression
- ✓ Simulation results
  - Real application
  - Asymptotic results
  - Summary

# Breast Cancer Expression Data

- Netherland Cancer Institute (van de Vijver et al. 2002)
- $n = 244$  breast cancer patients
- $p = 1217$  genes whose expression levels are significantly associated with the tumor progression ( $p$ -values from the univariate Cox model are smaller than  $5 \times 10^{-5}$ , with corresponding FDR  $< 0.01$ ).

# Hub Genes

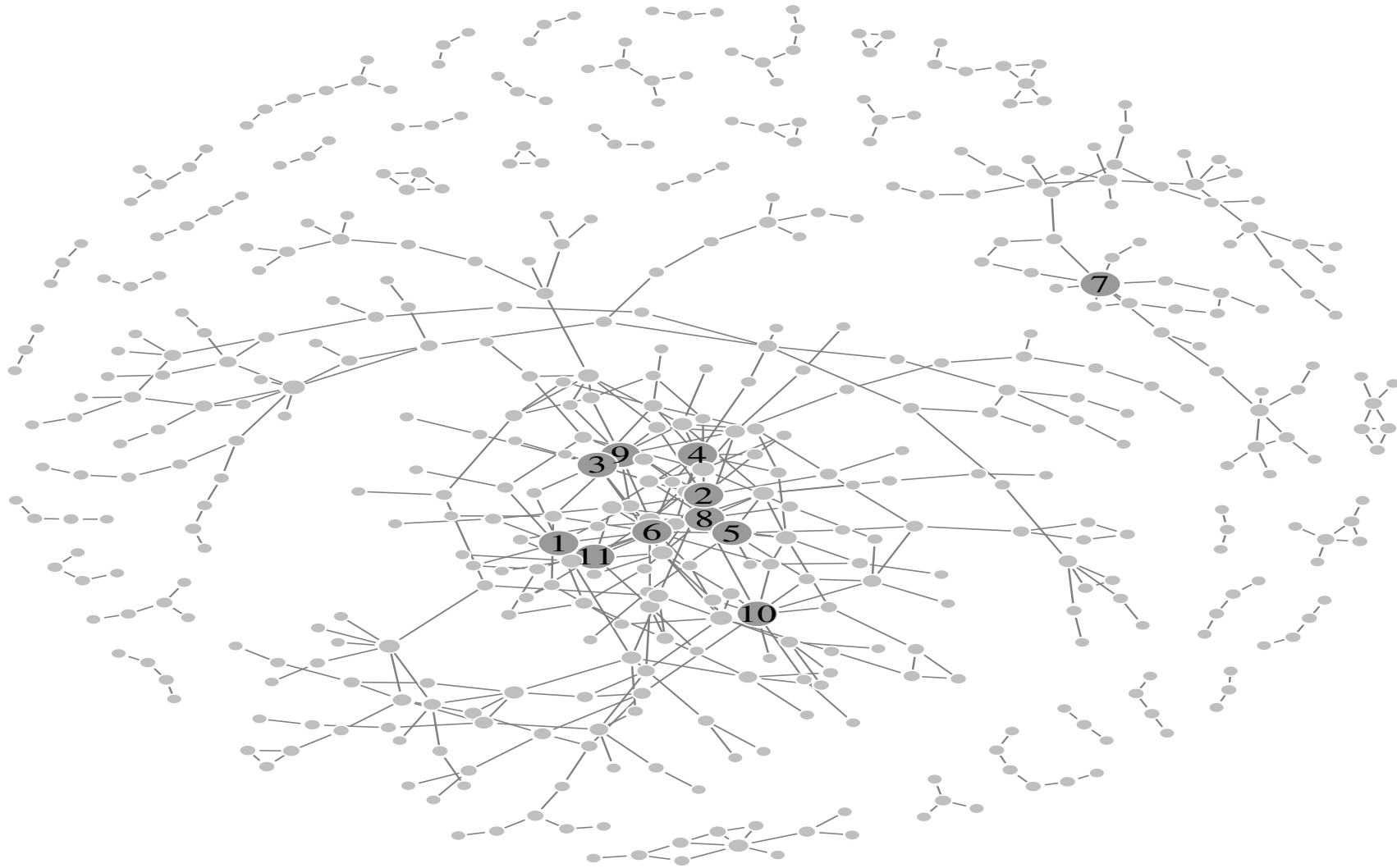


- There are 11 genes whose degree consistently rank the highest under various  $\lambda$ .
- 5 of the 11 candidate hub genes are important known regulator genes in breast cancer: HNF3A, KNSL6, STK12, RHD54L, BUB1.

## Inferred Network

When a total of 629 edges were selected

- 598 genes do not connect to any other genes;
- power law parameter was fitted to be  $\alpha = 2.56$ .



# Outline

- ✓ Background
- ✓ Joint sparse regression
- ✓ Simulation results
- ✓ Real application
  - Asymptotic results
  - Summary

# Asymptotic Results

- Let  $\boldsymbol{\rho}^*$  denote the true parameter vector.
- Let  $\mathcal{A} = \{(j, j') : \rho^{*jj'} \neq 0\}$  and  $q_n = |\mathcal{A}|$ .
- Under certain regularity conditions, if  $p = O(n^\kappa)$  for some  $\kappa > 0$ ,  $q_n \sim o(\sqrt{n/\log n})$  and the tuning parameter  $\lambda$  goes to zero at an appropriate rate, then we have
  - Estimation consistency:  $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*\| = o_p(1)$ .
  - Selection/sign consistency:

$$\Pr(\text{sign}(\hat{\rho}_{jj'}) = \text{sign}(\rho_{jj'}^*)) \rightarrow 1.$$

# Outline

- ✓ Background
- ✓ Joint sparse regression
- ✓ Simulation results
- ✓ Real application
- ✓ Asymptotic results
- Summary

# Summary

- We proposed a **joint** sparse regression model — **space**, which
  - controls the **overall** sparsity of the network;
  - is **flexible** for incorporating prior knowledge.
- We developed an **efficient algorithm**, active-shooting, for model fitting
  - Efficient enough for high-dimensional real applications and extensive simulation studies.
  - Can be naturally extended to other penalties.
- **space** outperforms two existing methods in both **edge detection** and **hub identification**.

- Real data application results in biologically sensible findings.
- The method is **consistent** (with diverging dimension) under suitable conditions.
- R package **space** is available on cran.