Methods for Robust High Dimensional Graphical Model Selection

Bala Rajaratnam

Department of Statistics
Stanford University

Fields Institute

Joint work with S. Oh (UC Berkeley) and K. Khare (U. Florida)

Motivation

- Availability of high-throughput data from various applications
- Need for methodology/tools for analyzing high-dimensional data
- Examples:
  - Biology: gene expression data
  - Environmental science: climate data on spatial grid
  - Finance: returns on thousands of stocks
  - Retail: consumer behavior
- Common goals:
  - Understand complex relationships & multivariate dependencies
  - Formulate correct models & develop inferential procedures

Modeling relationships

- Correlation: basic measure of linear pairwise relationships
- Covariance matrix Σ: collection of relationships
- Estimates of Σ required in procedures such as PCA, CCA, MANOVA, etc.
- Estimating (functions of) Σ and $\Omega = \Sigma^{-1}$ are of statistical interest
- Estimating Σ is difficult in high dimensions

Sparse estimates

- Matrix Σ or Ω of size p-by-p has $O(p^2)$ elements
- Estimating $O(p^2)$ parameters with classical estimators is not viable, especially when $n \ll p$
- **Reliably estimate small number of parameters** in Σ
- Model selection: zero/non-zero structure recovery
- Gives rise to sparse estimates of Σ or Ω
- Sparsity pattern can be represented by graphs/networks

Gaussian Graphical Models (GGM)

- Assume $Y = (Y_1, \ldots, Y_p)'$ has distribution $N_p(0, \Sigma)$
- Denote $V = \{1, 2, \ldots, p\}$
- Covariance matrix $cov(Y) = \Sigma$ encodes marginal dependencies

$$Y_i \perp\!\!\!\perp Y_j \iff cov(Y_i, Y_j) = [\Sigma]_{ij} = 0$$

- Inverse covariance matrix $\Omega = \Sigma^{-1}$ encodes conditional dependencies given the rest

$$\underbrace{\left(Y_i \perp\!\!\!\perp Y_j \mid Y_{V \setminus \{i,j\}}\right)}_{\text{conditional independence}} \iff \underbrace{[\Omega]_{ij} = 0}_{\text{matrix element}}$$

- Also known as Markov Random Fields (MRF)

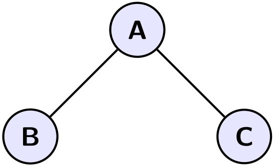
Gaussian Graphical Models (GGM)

- Graph summarizes relationships with nodes $V = \{1, \ldots, p\}$ and set E of edges

$$\underbrace{[\Omega]_{ij} = 0}_{\text{matrix element}} \iff \underbrace{i \not\sim j}_{\text{network/graph}}$$

- Build a graph from sparse Ω

$$\Omega = \begin{array}{cc} & \begin{array}{ccc} \mathsf{A} & \mathsf{B} & \mathsf{C} \end{array} \\ & \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 2 & 0 \\ 0.3 & 0 & 1.2 \end{pmatrix} \begin{array}{c} \mathsf{A} \\ \mathsf{B} \\ \mathsf{C} \end{array} \end{array}$$



GGM estimation algorithms

- **Regularized Gaussian likelihood methods**
  - Block coordinate descent (COVSEL) [Banerjee et al., 2008]
  - Graphical lasso (GLASSO) [Friedman, Hastie, & Tibshirani, 2008]
  - Large-scale GLASSO [Mazumder & Hastie, 2012]
  - QUIC [Hsieh et al., 2011]
  - G-ISTA [Guillot, Rajaratnam et al., 2012]
  - Graphical Dual Proximal Gradient Methods [Dalal & Rajaratnam, 2013]
  - Others
- **Bayesian methods**
  - Dawid & Lauritzen, 1993, Annals of Statistics
  - Letac & Massam, 2007, Annals of Statistics
  - Rajaratnam, Massam & Carvalho, 2008, Annals of Statistics
  - Khare & Rajaratnam, 2011, Annals of Statistics
  - Others
- **Testing-based methods**
  - Hero & Rajaratnam, 2011, JASA
  - Hero & Rajaratnam, 2012, IEEE, Information Theory

Regularized Gaussian likelihood graphical model selection

- All ℓ_1-regularized Gaussian-likelihood methods solve

$$\hat{\Omega} = \arg\max_{\Omega \succ 0} \{\log\det(\Omega) - \mathrm{tr}(\Omega S) - \lambda \|\Omega\|_1\}$$

- S: sample covariance matrix
- Graphical Lasso [Friedman, Hastie, & Tibshirani, 2008]
- Ω can be computed by solving optimization problem
- Adding ℓ_1-regularization term $\lambda\|\Omega\|_1$ introduces sparsity
- Penalty parameter λ controls level of sparsity
- Dependency on Gaussianity
  - Parametric model
  - Sensitivity to outliers
  - Log-concave function

Regularized pseudo-likelihood graphical model selection

- Two main main approaches:
  1. ℓ_1-regularized **likelihood** methods
  2. ℓ_1-regularized **regression-based/pseudo-likelihood** methods
- Series of linear regressions form a pseudo-likelihood function
- Objective function is the ℓ_1-penalized pseudo-likelihood
- Pseudo-likelihood assumes less about distribution of the data
- Applicable to wider range of data

Partial covariance and correlation

- Matrix $\mathbf{Y} \in \mathbb{R}^{n\times p}$ denotes iid observations of random variable with mean zero, covariance $\Sigma = \Omega^{-1}$.
- **Goal**: estimate partial correlation graph
- Partial correlation in terms of $\Omega = [\omega_{ij}]$:

$$\rho^{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

- Called "partial" because correlation of residuals r_k, where

$$r_k = \mathbf{Y}_k - \mathbf{Y}\hat{\beta}^{(k)}, \text{ where } \hat{\beta}^{(k)} = \arg\min_{\beta:\beta_k=0} \left\{\|\mathbf{Y}_k - \mathbf{Y}\beta\|_2^2\right\}.$$

  Now, partial correlation is $\rho^{ij} = cor(r_i, r_j)$
- It can be shown that $\underbrace{\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}}_{\beta_{ij}}$
- Zero/non-zero pattern of $[\rho^{ij}]$ is identical to that of Ω
- Partial correlation graph is given by sparsity pattern of Ω

Regularized regression-based graphical model selection

- Neighborhood selection (NS)[Meinshausen and Bühlmann, 2006]

$$\hat{\omega}^{(i)} = \arg\min_{\beta:\beta_i=0} \left\{\|\mathbf{Y}_i - \mathbf{Y}\beta\|_2^2 + \lambda\|\beta\|_1\right\}$$

- Neighborhood of i is defined as

$$\widehat{\mathrm{ne}}^{(i)} = \{k : \hat{\omega}_k^{(i)} \neq 0\}$$

- MB does not take into account symmetry of Ω

$$j \in \widehat{\mathrm{ne}}^{(i)} \quad \not\Longrightarrow \quad i \in \widehat{\mathrm{ne}}^{(j)}$$

- Current state-of-the-art methods address this issue
  - SPACE [Peng et al., 2009]
  - SYMLASSO [Friedman, Hastie, & Tibshirani, 2010]
  - SPLICE [Rocha et al., 2008]

Sparse PArtial Correlation Estimation [Peng et al., 2009]

SPACE objective function: $(w_i = 1 \text{ or } w_i = \omega_{ii})$

$$Q_{\mathrm{spc}}(\Omega) =: -\frac{1}{2}\sum_{i=1}^{p} n \log \omega_{ii} + \frac{1}{2}\sum_{i=1}^{p} w_i \|\mathbf{Y}_i - \sum_{j\neq i} \underbrace{\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}}_{\beta_{ij}} \mathbf{Y}_j\|_2^2 + \lambda \sum_{1\leqslant i<j\leqslant p} |\rho^{ij}|$$

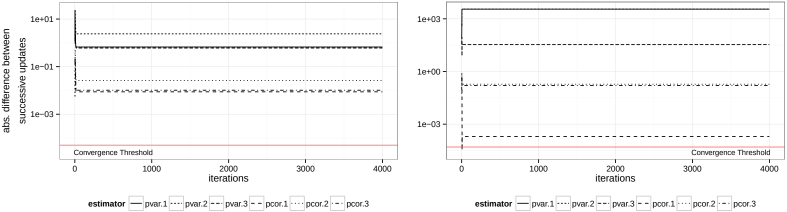
1. **Update** $[\rho^{ij}]$ coordinate-wise (using current estimates $[\hat{\omega}_{ii}]$):

$$[\rho^{ij}] \leftarrow \min_{[\rho^{ij}]} \left\{ \frac{1}{2}\sum_{i=1}^{p} w_i \|\mathbf{Y}_i - \sum_{j\neq i} \rho^{ij}\sqrt{\frac{\hat{\omega}_{jj}}{\hat{\omega}_{ii}}} \mathbf{Y}_j\|_2^2 + \lambda \sum_{1\leqslant i<j\leqslant p} |\rho^{ij}| \right\}$$

2. **Update** $[\omega_{ii}]$ (using current estimates $[\hat{\rho}^{ij}]$ and $[\hat{\omega}_{ii}]$):

$$\omega_{ii} \leftarrow \left(\|\mathbf{Y}_i - \sum_{j\neq i} \hat{\rho}^{ij}\sqrt{\frac{\hat{\omega}_{jj}}{\hat{\omega}_{ii}}} \mathbf{Y}_j\|_2^2 \right)^{-1}$$

Non-converging example: $p = 3$ case



(a) SPACE ($w_i = \omega_{ii}$)

(b) SPACE ($w_i = 1$)

Figure: $\mathbf{Y}^{(\mathbf{i})} \sim \mathcal{N}_3(0, \Omega^{-1})$, (left) $n = 4$, (right) $n = 100$

Non-convergence of SPACE

- Investigate the nature and extent of convergence issues:
  1. Are such examples pathological? How widespread are they?
  2. When do they occur ?
- Consider a sparse 100×100 matrix Ω with edge density 4% and condition number of 100.
- Generate 100 multivariate Gaussian datasets (with $n = 100$), $\mu = 0$ and $\Sigma = \Omega^{-1}$.
- Record the number of times (out of 100) for which SPACE1 (uniform weights) and SPACE2 (partial variance weights) do not converge within 1500 iterations.
- Original implementation of SPACE by [Peng et al., 2009] claims only 3 iterations are sufficient.

Non-convergence of SPACE

SPACE1 ($w_i = 1$)			SPACE2 ($w_i = \omega_{ii}$)		
λ^*	NZ	NC	λ^*	NZ	NC
0.026	60.9%	92	0.085	79.8%	100
0.099	19.7%	100	0.160	28.3%	0
0.163	7.6%	100	0.220	10.7%	0
0.228	2.9%	100	0.280	4.8%	0
0.614	0.4%	0	0.730	0.5%	97

Table: Number of simulations (out of 100) that do not converge within 1500 iterations (NC) for select values of penalty parameter ($\lambda^* = \lambda/n$). Average percentage of non-zeros (NZ) in $\hat{\Omega}$ are also shown.

- SPACE exhibits extensive non-convergence behavior
- Problem exacerbated when condition number is high
- Typical of high dimensional sample starved settings

Symmetric Lasso and SPLICE

SYMLASSO [Friedman, Hastie, & Tibshirani, 2010]:

$$Q_{\mathrm{sym}}(\boldsymbol{\alpha}, \breve{\Omega}) = \frac{1}{2} \sum_{i=1}^{p} \left[n \log \alpha_{ii} + \frac{1}{\alpha_{ii}} \|\mathbf{Y}_i + \sum_{j \neq i} \omega_{ij} \alpha_{ii} \mathbf{Y}_j\|^2 \right] + \lambda \sum_{1 \leqslant i < j \leqslant p} |\omega_{ij}|,$$

where $\alpha_{ii} = 1/\omega_{ii}$.

SPLICE [Rocha et al., 2008]:

$$Q_{\mathrm{spl}}(\mathbf{B}, \mathbf{D}) = \frac{n}{2} \sum_{i=1}^{p} \log(d_{ii}^2) + \frac{1}{2} \sum_{i=1}^{p} \frac{1}{d_{ii}^2} \|\mathbf{Y}_i - \sum_{j \neq i} \beta_{ij} \mathbf{Y}_j\|^2 + \lambda \sum_{i<j} |\beta_{ij}|,$$

where $d_{ii}^2 = \omega_{ii}$.

Also, alternating (off-diagonal vs diagonal) iterative algorithms
No convergence guarantees

Regularized regression-based graphical model selection

- **Advantage**: Regression-based methods perform better model selection than likelihood-based methods in finite sample
- **Advantage**: Regression-based methods are less restrictive than Gaussian likelihood-based methods
- **Disadvantage**: $\hat{\Omega}$ may not be positive definite (can be fixed)
- **Disadvantage**: Solution may not be computable
- **Cause**: Iterative algorithms SPACE, SYMLASSO and SPLICE are not guaranteed to converge

Regression-based methods: summary

	METHOD			
Property	NS	SPACE	SYMLASSO	SPLICE
Symmetry		$+$	$+$	$+$
Convergence guarantee	N/A			
Asymptotic consistency ($n, p \to \infty$)	$+$	$+$		

How can we obtain all of the good properties simultaneously?

Design goals of a new pseudo-likelihood approach

- Can we design a regression-based approach that guarantees existence of a solution?
- Is there a better chance of guaranteeing a well defined solution if a convex formulation is developed?
- Advantages of a convex formulation:
  - Easier analysis of theoretical properties
  - Better chance of algorithmic convergence
  - Global minimum is guaranteed to exist
- Can we leverage convex optimization theory?
- Current pseudo-likelihood methods are not jointly convex (in the parametrization proposed in the respective papers)
- Can we develop a convex formulation of pseudo-likelihood graphical model selection problem?

Convex formulation of graphical model selection problem

Revisit the SPACE objective function

$$Q_{\mathrm{spc}}(\Omega) =: -\frac{1}{2}\sum_{i=1}^{p} n\log\omega_{ii} + \frac{1}{2}\sum_{i=1}^{p} w_i \|\mathbf{Y}_i - \sum_{j\neq i} \underbrace{\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}}_{\beta_{ij}} \mathbf{Y}_j\|_2^2 + \lambda \sum_{1\leqslant i<j\leqslant p} |\rho^{ij}|$$

- $Q_{\mathrm{spc}}(\Omega)$ is not jointly convex in elements of Ω
- Since $\beta_{ij} = \rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}} = -\frac{\omega_{ij}}{\omega_{ii}}$, regression term is not convex
- Since $|\rho^{ij}| = \left|-\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}\right|$, penalty term is not convex

Convex formulation of graphical model selection problem

Consider,

$$w_i\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}\mathbf{Y}_j\|_2^2 = w_i\|\mathbf{Y}_i + \sum_{j\neq i}\frac{\omega_{ij}}{\omega_{ii}}\mathbf{Y}_j\|_2^2 \qquad \left(\because \rho^{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}\right)$$

$$= w_i\|\frac{1}{\omega_{ii}}(\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j)\|_2^2$$

$$= \frac{w_i}{\omega_{ii}^2}\|\sum_{j=1}^{p}\omega_{ij}\mathbf{Y}_j\|_2^2$$

Now, let $w_i = \omega_{ii}^2$, then

$$\|\sum_{j=1}^{p}\omega_{ij}\mathbf{Y}_j\|_2^2 = \omega_{\bullet i}'\mathbf{Y}'\mathbf{Y}\omega_{\bullet i} \geqslant 0, \text{ (quadratic form)}$$

Therefore, Q_{con} below is jointly convex:

$$Q_{\text{con}}(\Omega) =: -\sum_{i=1}^{p} n\log\omega_{ii} + \frac{1}{2}\sum_{i=1}^{p}\|\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j\|_2^2 + \lambda\sum_{1\leqslant i<j\leqslant p}|\omega_{ij}|$$

Establishing properties of CONCORD

Optimization properties

- Task 1: (**Optimization algorithm**) Can we find an effective algorithm to minimize the $Q_{\text{con}}(\Omega)$ so that a solution always exists and is computable?
- Task 2: (**Guarantee of convergence to global optimum**) Can we establish convergence? Do we have a globally optimal solution?
- Task 3: (**Computational complexity**) What is the computational complexity of the optimization method? Is it competitive with other methods?
- Task 4: (**Running time comparison**) How do the actual running times compare with other methods?

Establishing properties of CONCORD

Statistical Properties

- Task 5: (**Consistency and Large Sample properties**) Are Concord estimates guaranteed to recover the true underlying partial correlation graphs for data generated from such models?
- Task 6: (**Finite sample properties**) How does CONCORD perform in terms of recovering the partial correlation graph in finite sample settings?
- Task 7: (**Applications**) How does CONCORD perform in applications in comparison with other methods where high dimensional covariance estimates are required?

Goal: To investigate the above questions systematically

CONvex CORrelation selection methoD (CONCORD)

CONCORD objective function:

$$Q_{\text{con}}(\Omega) =: -\sum_{i=1}^{p} n \log \omega_{ii} + \frac{1}{2}\sum_{i=1}^{p} \|\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i} \omega_{ij}\mathbf{Y}_j\|_2^2 + \lambda \sum_{1\leqslant i<j\leqslant p} |\omega_{ij}|$$

Coordinate-wise iterative algorithm

1. **Update** $[\omega_{ij}]$[1] (other coefficients held constant):

$$\omega_{ij} \leftarrow \frac{S_{\frac{\lambda}{n}}\left(-\left(\sum_{j'\neq j}\omega_{ij'}S_{jj'} + \sum_{i'\neq i}\omega_{i'j}S_{ii'}\right)\right)}{S_{ii}+S_{jj}}$$

2. **Update** $[\omega_{ii}]$ (other coefficients held constant):

$$\omega_{ii} \leftarrow \frac{-\sum_{j\neq i}\omega_{ij}S_{ij} + \sqrt{\left(\sum_{j\neq i}\omega_{ij}S_{ij}\right)^2 + 4S_{ii}}}{2S_{ii}}$$

[1] Soft-thresholding operator: $S_\lambda(x) = \text{sign}(x)(|x|-\lambda)_+$

Optimization aspects of CONCORD algorithm

Theorem: Let $\mathcal{A}_p$ denote space of $p \times p$ symmetric matrices. Also, let $\mathcal{M} \subset \mathcal{A}_p$ denote a subspace such that

$$\mathcal{M} := \{\Omega \in \mathcal{A}_p : \omega_{ii} > 0, \text{ for every } 1 \leqslant i \leqslant p\}.$$

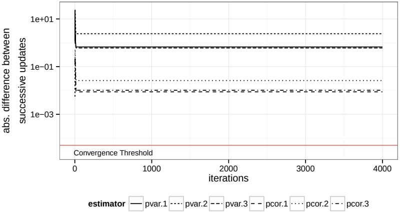
If $Y_i \neq 0$ for every $1 \leqslant i \leqslant p$, the sequence of iterates $\{\hat{\Omega}^{(r)}\}_{r \geqslant 0}$ obtained by the CONCORD algorithm converges to a global minimum of $Q_{\text{con}}(\Omega)$. More specifically,

$$\hat{\Omega}^{(r)} \to \hat{\Omega} \in \mathcal{M} \text{ as } r \to \infty$$

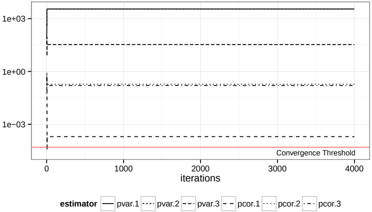
for some $\hat{\Omega}$, and furthermore

$$Q_{\text{con}}(\hat{\Omega}) \leqslant Q(\Omega) \text{ for all } \Omega \in \mathcal{M}.$$

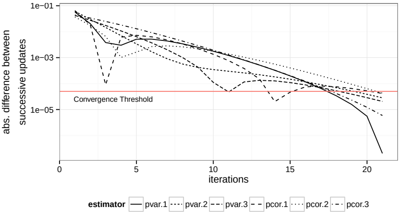
Non-converging example: $p = 3$ case



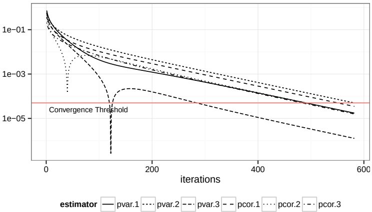
(a) SPACE ($w_i = \omega_{ii}$)



(b) SPACE ($w_i = 1$)



(c) CONCORD



(d) CONCORD

Figure: $\mathbf{Y}^{(\mathbf{i})} \sim \mathcal{N}_3(0, \Omega^{-1})$, (left) $n = 4$, (right) $n = 100$

Computational complexity of CONCORD algorithm

- GLASSO: $O(p^3)$
- SPACE: $\min(O(np^2), O(p^3))$
- SYMLASSO: $\min(O(np^2), O(p^3))$
- CONCORD: $\min(O(np^2), O(p^3))$

Running time of CONCORD: I

$\mathbf{p} = \mathbf{1000}$, $\mathbf{n} = \mathbf{200}$					
GLASSO			CONCORD		
λ	NZ	Time	λ^*	NZ	Time
0.14	4.77%	87.60	0.12	4.23%	6.12
0.19	0.87%	71.47	0.17	0.98%	5.10
0.28	0.17%	5.41	0.28	0.15%	5.37
0.39	0.08%	5.30	0.39	0.07%	4.00
0.51	0.04%	6.38	0.51	0.04%	4.76

$\mathbf{p} = \mathbf{1000}$, $\mathbf{n} = \mathbf{200}$					
SPACE1 ($w_i = 1$)			SPACE2 ($w_i = \omega_{ii}$)		
λ	NZ	Time	λ^*	NZ	Time
0.10	4.49%	101.78	0.16	100.00%	19206.55
0.17	0.64%	99.20	0.21	1.76%	222.00
0.28	0.14%	138.01	0.30	0.17%	94.59
0.39	0.07%	75.55	0.40	0.08%	108.61
0.51	0.04%	49.59	0.51	0.04%	132.34

Table: Timing comparison (seconds) for $p = 1000$, λ = penalty parameter, $\lambda^* = \lambda/n$ for CONCORD/SPACE. NZ = the percentage of non-zero entries

Running time of CONCORD: II

p = 3000, n = 600					
GLASSO			CONCORD		
λ	NZ	Time	λ^*	NZ	Time
0.09	2.71%	1842.74	0.09	2.10%	266.69
0.10	1.97%	1835.32	0.10	1.59%	235.49
0.10	1.43%	1419.41	0.10	1.19%	232.67

p = 3000, n = 900					
GLASSO			CONCORD		
λ	NZ	Time	λ^*	NZ	Time
0.09	0.70%	1389.96	0.09	0.64%	298.21
0.10	0.44%	1395.42	0.10	0.41%	298.00
0.10	0.27%	1334.78	0.10	0.26%	302.15

Table: Timing comparison (seconds) for $p = 3000$, λ = penalty parameter, $\lambda^* = \lambda/n$ for CONCORD. NZ = the percentage of non-zero entries

- CONCORD is highly competitive.
- Orders of magnitude faster in high dimensional settings.
- SPACE is slow to converge when $n \ll p$.

Large sample properties: Assumptions

For sample size n and number of feature $p = p_n$, assume
True inverse covariance matrix: $\bar{\Omega}_n = [\bar{\omega}_{n,ij}]$, $1 \leqslant i, j \leqslant p_n$, and
$\bar{\omega}_n^o$ denotes the off-diagonal elements.
Assumptions:

- A0: Accurate estimates of diagonals $\hat{\alpha}_{n,ii}$:

$$\max_{1 \leqslant i \leqslant p_n} |\widehat{\alpha}_{n,ii} - \bar{\omega}_{ii}| \leqslant C \left(\sqrt{\frac{\log n}{n}} \right),$$

  holds with probability larger than $1 - O(n^{-\eta})$.

- A1: Bounded eigenvalues: eigenvalues of $\bar{\Omega}_n$ are such that

$$\lambda_{min} > 0 \text{ and } \lambda_{max} < \infty \text{ , for all } n$$

- A2: Sub-Gaussianity,

- A3: Incoherence condition

Large sample properties: Theorem

Suppose that assumptions (A0)-(A3) are satisfied. Suppose $p_n = O(n^\kappa)$ for some $\kappa > 0$, $q_n = o\left(\sqrt{n}\log n\right)$, $\sqrt{\frac{q_n \log n}{n}} = o(\lambda_n)$, $\lambda_n \sqrt{n} \log n \to \infty$, and $\sqrt{q_n}\lambda_n \to 0$, as $n \to \infty$.

Then there exists a constant C such that for any $\eta > 0$, the following events hold with probability at least $1 - O(n^{-\eta})$.

- There exists a minimizer $\widehat{\omega}_n^o = ((\widehat{\omega}_{n,ij}))_{1 \leqslant i < j \leqslant p_n}$ of $Q_{\text{con}}(\omega^o, \widehat{\boldsymbol{\alpha}}_n)$.
- Any minimizer $\widehat{\omega}_n^o$ of $Q_{\text{con}}(\omega^o, \widehat{\boldsymbol{\alpha}}_n)$ satisfies

$$\|\widehat{\omega}_n^o - \bar{\omega}_n^o\|_2 \leqslant C\sqrt{q_n}\lambda_n \quad \textbf{(Parameter consistency)}$$

and

$$\text{sign}(\widehat{\omega}_{n,ij}) = \text{sign}(\bar{\omega}_{n,ij}),\ \forall\ 1 \leqslant i < j \leqslant p_n \quad \textbf{(Sign consistency)}.$$

Model selection in finite samples

- A $p \times p$ sparse positive definite matrix Ω (with $p = 1000$) with condition number 13.6.
- Sample size $n = 200, 400, 800$, 50 datasets, each having *i.i.d.* multivariate-t distribution with $\mu = 0$, $\Sigma = \Omega^{-1}$.
- Compare model selection performance: area-under-the-curve (AUC) of ROC curves

	n = 200		**n = 400**		**n = 800**	
Solver	Median	IQR	Median	IQR	Median	IQR
GLASSO	0.745	0.032	0.819	0.030	0.885	0.029
CONCORD	0.811	0.011	0.887	0.012	0.933	0.013

Table: Median and IQR of AUC for 50 simulations.

- CONCORD has a higher AUC for each of the 150 datasets.
- CONCORD not only recovers the sparsity structure more accurately, it also has much less variation.

CONCORD method: summary

	METHOD				
Property	NS	SPACE	SYMLASSO	SPLICE	CONCORD
Symmetry		$+$	$+$	$+$	$+$
Convergence guarantee (fixed n)	N/A				$+$
Asymptotic consistency ($n, p \to \infty$)	$+$	$+$			$+$

Yes! CONCORD retains all good properties

CONCORD method: summary

- Optimization aspects
  - **Jointly convex formulation**
  - Theoretical **guarantee of convergence**
  - Converges to **globally optimal solution**
- Statistical properties
  - **Asymptotically consistent** estimator as $n, p \to \infty$
  - **Competitive with other pseudo-likelihood methods** in finite sample
- Computational cost
  - **Computationally complexity is competitive**

Unifying framework for regression-based/pseudo-likelihood graphical model selection

What are we solving exactly?

$$\mathcal{L}_{\text{con}}(\Omega) = \frac{1}{2}\sum_{i=1}^{p}\left[-n\log\omega_{ii}^2 + \quad \|\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j\|_2^2\right]$$

$$\mathcal{L}_{\text{spc},1}(\Omega_D, \boldsymbol{\rho}) = \frac{1}{2}\sum_{i=1}^{p}\left[-n\log\omega_{ii} + \quad \|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}\mathbf{Y}_j\|_2^2\right]$$

$$\mathcal{L}_{\text{spc},2}(\Omega_D, \boldsymbol{\rho}) = \frac{1}{2}\sum_{i=1}^{p}\left[-n\log\omega_{ii} + \omega_{ii}\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}\mathbf{Y}_j\|_2^2\right]$$

$$\mathcal{L}_{\text{sym}}(\boldsymbol{\alpha}, \Omega_F) = \frac{1}{2}\sum_{i=1}^{p}\left[\quad n\log\alpha_{ii} + \frac{1}{\alpha_{ii}}\|\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\alpha_{ii}\mathbf{Y}_j\|^2\right]$$

$$\mathcal{L}_{\text{spl}}(\mathbf{B}, \mathbf{D}) = \frac{1}{2}\sum_{i=1}^{p}\left[\quad n\log(d_{ii}^2) + \frac{1}{d_{ii}^2}\|\mathbf{Y}_i - \sum_{j\neq i}\beta_{ij}\mathbf{Y}_j\|_2^2\right],$$

Unifying framework lemma: part 1

The above pseudo-likelihoods (up to reparameterization) can be expressed in matrix form as follows:

$$\mathcal{L}_{\text{con}}(\Omega) = \frac{n}{2}\left[-\log\det\Omega_D^2 + \text{tr}(\mathbf{S}\Omega^2)\right]$$
$$\mathcal{L}_{\text{spc},1}(\Omega) = \frac{n}{2}\left[-\log\det\Omega_D + \text{tr}(\mathbf{S}\Omega\Omega_D^{-2}\Omega)\right]$$
$$\mathcal{L}_{\text{spc},2}(\Omega) = \frac{n}{2}\left[-\log\det\Omega_D + \text{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right]$$
$$\mathcal{L}_{\text{sym}}(\Omega) = \frac{n}{2}\left[-\log\det\Omega_D + \text{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right]$$
$$\mathcal{L}_{\text{spl}}(\Omega) = \frac{n}{2}\left[-\log\det\Omega_D + \text{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right],$$

where $\Omega_D = \text{diag}(\Omega)$

Unifying framework lemma: part 2

Generalized form of pseudo-log-likelihood

$$\mathcal{L}_{\text{uni}}(G(\Omega), H(\Omega)) = \frac{n}{2}\left[-\log\det G(\Omega) + \operatorname{tr}(\mathbf{S}H(\Omega))\right],$$

where $G(\Omega)$ and $H(\Omega)$ are functions of Ω.

Standard Gaussian log-likelihood when $G(\Omega) = H(\Omega) = \Omega$:

$$\mathcal{L}_{\text{Gaussian}}(\Omega) = \mathcal{L}_{\text{uni}}(\Omega, \Omega) = \frac{n}{2}\left[-\log\det\Omega + \operatorname{tr}(\mathbf{S}\Omega)\right]$$

Insights for SPACE2, SYMLASSO and SPLICE

SPACE2, SYMLASSO and SPLICE formulations:

$$\mathcal{L}_{\mathsf{uni}}(\Omega_D, \Omega\Omega_D^{-1}\Omega) = \frac{n}{2}\left[-\log\det\Omega_D + \mathrm{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right]$$

- Three of the four pseudo-likelihoods are equivalent up to reparameterizations
- Three methods apply different ℓ_1-penalties

Applications of graphical model selection and (inverse) covariance estimation

Biological application: gene co-expression of breast cancer

- Breast cancer gene expression study [Cheng et al., 2009]
- $n = 248$ and other clinical data (metastasis, tumor size, etc..)
- Reduce to $\sim$1100 genes by survival analysis (from $\sim$20000)
- Select λ such that 200 non-zero elements remain in $\hat{\Omega}$
- Identify most highly connected (hub) genes [Carter et al., 2004, Jeong et al., 2001, Han et al., 2004]

Biological application: gene co-expression of breast cancer

Gene Symbol	CONCORD	SYMLASSO	SPACE1	SPACE2	Reference
HNF3A (FOXA1)	+	+	+	+	[Koboldt and Others, 2012, Albergaria et al., 2009, Davidson et al., 2011, Lacroix and Leclercq, 2004, Robinson et al., 2011]
TONDU	+	+	+	+	
FZD9	+	+	+	+	[Katoh, 2008, Rø nneberg et al., 2011]
KIAA0481	+	+	+	+	[Gene record discontinued]
KRT16	+	+	+		[Glinsky et al., 2005, Joosse et al., 2012, Pellegrino et al., 1988]
KNSL6 (KIF2C)	+			+	[Eschenbrenner et al., 2011, Shimo et al., 2007, Shimo et al., 2008]
FOXC1	+	+	+	+	[Du et al., 2012, Sizemore and Keri, 2012, Wang et al., 2012, Ray et al., 2011, Tkocz et al., 2012]
PSA	+	+		+	[Kraus et al., 2010, Mohajeri et al., 2011, Sauter et al., 2004, Yang et al., 2002]
GATA3	+	+	+	+	[Koboldt and Others, 2012, Davidson et al., 2011, Albergaria et al., 2009, Eeckhoute et al., 2007, Jiang et al., 2010, Licata et al., 2010, Yan et al., 2010]
C20ORF1 (TPX2)	+				[Maxwell and Others, 2011, Bibby et al., 2009]
E48		+	+	+	
ESR1				+	[Zheng et al., 2012]

[Maxwell and Others, 2011] identifies a regulatory mechanism involving TPX2, Aurora A, RHAMM and BRCA1 genes in breast cancer

TPX2 gene in breast cancer

- [Maxwell and Others, 2011] is an extensive study involving thousands of breast cancer patients
- Breast cancer type 1 susceptibility protein (BRCA1), a known gene related to breast cancer
- TPX2 gene is identified as having strong link to BRCA1
- **“Reorganization (of microtubules) is facilitated by BRCA1 and impaired by AURKA, which is regulated by negative feedback involving RHAMM and TPX2.”** [Maxwell et al., 2011]

Financial application: portfolio optimization

Dow-Jones Index:

- Index of 30 stocks
- Mean-variance portfolio (MVP) theory uses covariance matrix to hedge risk
- Simplest variant: minimum variance portfolio (given Σ)

$$\begin{array}{ll} \text{minimize} & w^T \Sigma w \\ \text{subject to} & \mathbf{1}^T w = 1 \end{array}$$

  Analytical solution: $w^\star = (\mathbf{1}^T \Sigma^{-1} \mathbf{1})^{-1} \Sigma^{-1} \mathbf{1}$

- Due to non-stationarity, use rebalancing strategy:
  Every 4 weeks, use past N_{est} days for $\hat{\Sigma} = \hat{\Omega}^{-1}$

Financial application: portfolio optimization

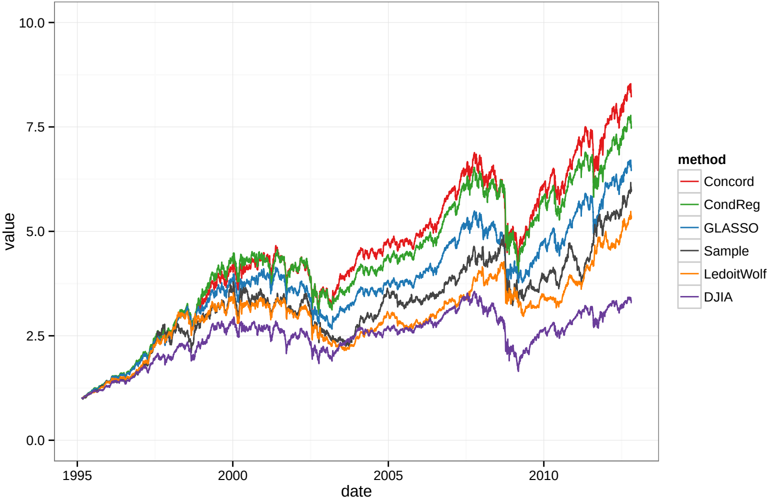


Figure: $N_{\mathrm{est}} = 75$ days, rebalance every 4 weeks

Finance: Minimum variance portfolio returns

Return measure: mean excess return per unit of risk

$$\textbf{Sharpe ratio} = \frac{\mathbb{E}(R_t - R_f)}{\sqrt{\mathrm{Var}\,(R_t)}}, \text{ where } R_f = 3\% \text{ (annual) is chosen}$$

N_{est}	DJIA	Sample	GLASSO	Concord	CondReg	LedoitWolf
35	2.09	2.77	4.01	**4.12**	4.06	**4.10**
40	2.09	3.44	3.93	**4.10**	**3.98**	3.91
45	2.09	2.43	3.78	**3.98**	**3.85**	3.59
50	2.09	2.31	3.81	**4.06**	**3.89**	3.71
75	2.09	3.40	3.70	**4.04**	**3.89**	3.49
	References		Sparse models		Dense estimates	

Table: Penalty λ chosen with cross-validation to minimize RSS, (values multiplied by 100)

Applications: summary

- Biological example: hub gene discovery
  - Discovered empirically validated genes
  - Other methods are useful too!
- Finance example: minimum variance portfolio selection
  - CONCORD estimator yields best Sharpe ratio even better than Ledoit-Wolf
  - Graphical model selection methods adapt to changing covariance structure

Thank you!

References I

Albergaria, A., Paredes, J., Sousa, B., Milanezi, F., Carneiro, V., Bastos, J., Costa, S., Vieira, D., Lopes, N., Lam, E. W., Lunet, N., and Schmitt, F. (2009).
Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours.
Breast Cancer Research, 11(3):R40.

Banerjee, O., El Ghaoui, L., and D'Aspremont, A. (2008).
Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data.
The Journal of Machine Learning Research, 9:485–516.

Bibby, R. A., Tang, C., Faisal, A., Drosopoulos, K., Lubbe, S., Houlston, R., Bayliss, R., and Linardopoulos, S. (2009).
A cancer-associated aurora A mutant is mislocalized and misregulated due to loss of interaction with TPX2.
The Journal of Biological Chemistry, 284(48):33177–84.

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004).
Gene co-expression network topology provides a framework for molecular characterization of cellular state.
Bioinformatics (Oxford, England), 20(14):2242–50.

Cheng, C.-J., Lin, Y.-C., Tsai, M.-T., Chen, C.-S., Hsieh, M.-C., Chen, C.-L., and Yang, R.-B. (2009).
SCUBE2 Suppresses Breast Tumor Cell Proliferation and Confers a Favorable Prognosis in Invasive Breast Cancer.
Cancer Research, 69(8):3634–3641.

Davidson, B., Stavnes, H. T., Holth, A., Chen, X., Yang, Y., Shih, I.-M., and Wang, T.-L. (2011).
Gene expression signatures differentiate ovarian/peritoneal serous carcinoma from breast carcinoma in effusions.
Journal of Cellular and Molecular Medicine, 15(3):535–44.

Du, J., Li, L., Ou, Z., Kong, C., Zhang, Y., Dong, Z., Zhu, S., Jiang, H., Shao, Z., Huang, B., and Lu, J. (2012).
FOXC1, a target of polycomb, inhibits metastasis of breast cancer cells.
Breast Cancer Research and Treatment, 131(1):65–73.

References II

Eeckhoute, J., Keeton, E. K., Lupien, M., Krum, S. A., Carroll, J. S., and Brown, M. (2007).
Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer.
Cancer Research, 67(13):6477–83.

Eschenbrenner, J., Winsel, S., Hammer, S., Sommer, A., Mittelstaedt, K., Drosch, M., Klar, U., Sachse, C., Hannus, M., Seidel, M., Weiss, B., Merz, C., Siemeister, G., and Hoffmann, J. (2011).
Evaluation of activity and combination strategies with the microtubule-targeting drug sagopilone in breast cancer cell lines.
Frontiers in Oncology, 1:44.

Friedman, J., Hastie, T., and Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441.

Glinsky, G. V., Berezovska, O., and Glinskii, A. B. (2005).
Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer.
The Journal of clinical investigation, 115(6):1503–21.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and Vidal, M. (2004).
Evidence for dynamically organized modularity in the yeast protein-protein interaction network.
Nature, 430(6995):88–93.

Jeong, H., Mason, S. P., Barabasi, A.-L., and Oltvai, Z. N. (2001).
Lethality and centrality in protein networks.
Nature, 411(6833):41–42.

Jiang, S., Katayama, H., Wang, J., Li, S. A., Hong, Y., Radvanyi, L., Li, J. J., and Sen, S. (2010).
Estrogen-induced aurora kinase-A (AURKA) gene expression is activated by GATA-3 in estrogen receptor-positive breast cancer cells.
Hormones & Cancer, 1(1):11–20.

References III

Joosse, S. A., Hannemann, J., Spötter, J., Bauche, A., Andreas, A., Müller, V., and Pantel, K. (2012).
Changes in Keratin Expression during Metastatic Progression of Breast Cancer: Impact on the Detection of Circulating Tumor Cells.
Clinical cancer research : an official journal of the American Association for Cancer Research, 18(4):993–1003.

Katoh, M. (2008).
WNT signaling in stem cell biology and regenerative medicine.
Current Drug Targets, 9(7):565–70.

Koboldt, D. C. and Others (2012).
Comprehensive molecular portraits of human breast tumours.
Nature, 490(7418):61–70.

Kraus, T. S., Cohen, C., and Siddiqui, M. T. (2010).
Prostate-specific antigen and hormone receptor expression in male and female breast carcinoma.
Diagnostic Pathology, 5:63.

Lacroix, M. and Leclercq, G. (2004).
About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer.
Molecular and Cellular Endocrinology, 219(1-2):1–7.

Licata, L. A., Hostetter, C. L., Crismale, J., Sheth, A., and Keen, J. C. (2010).
The RNA-binding protein HuR regulates GATA3 mRNA stability in human breast cancer cell lines.
Breast Cancer Research and Treatment, 122(1):55–63.

Maxwell, C. A. and Others (2011).
Interplay between BRCA1 and RHAMM regulates epithelial apicobasal polarization and may influence risk of breast cancer.
PLoS Biology, 9(11):e1001199.

References IV

Mazumder, R. and Hastie, T. (2012).
Exact Covariance Thresholding into Connected Components for Large-Scale Graphical Lasso.
The Journal of Machine Learning Research, 13:781–794.

Meinshausen, N. and Bühlmann, P. (2006).
High-dimensional graphs and variable selection with the Lasso.
The Annals of Statistics, 34(3):1436–1462.

Mohajeri, A., Zarghami, N., Pourhasan Moghadam, M., Alani, B., Montazeri, V., Baiat, A., and Fekhrjou, A. (2011).
Prostate-specific antigen gene expression and telomerase activity in breast cancer patients: possible relationship to steroid hormone receptors.
Oncology Research, 19(8-9):375–80.

Pellegrino, M. B., Asch, B. B., Connolly, J. L., and Asch, H. L. (1988).
Differential expression of keratins 13 and 16 in normal epithelium, benign lesions, and ductal carcinomas of the human breast determined by the monoclonal antibody Ks8.12.
Cancer Research, 48(20):5831–6.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009).
Partial Correlation Estimation by Joint Sparse Regression Models.
Journal of the American Statistical Association, 104(486):735–746.

Ray, P. S., Bagaria, S. P., Wang, J., Shamonki, J. M., Ye, X., Sim, M.-S., Steen, S., Qu, Y., Cui, X., and Giuliano, A. E. (2011).
Basal-like breast cancer defined by FOXC1 expression offers superior prognostic value: a retrospective immunohistochemical study.
Annals of Surgical Oncology, 18(13):3839–47.

References V

Rø nneberg, J. A., Fleischer, T., Solvang, H. K., Nordgard, S. H., Edvardsen, H., Potapenko, I., Nebdal, D., Daviaud, C., Gut, I., Bukholm, I., Naume, B. r., Bø rresen Dale, A.-L., Tost, J., and Kristensen, V. (2011).
Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer.
Molecular Oncology, 5(1):61–76.

Robinson, J. L. L., Macarthur, S., Ross-Innes, C. S., Tilley, W. D., Neal, D. E., Mills, I. G., and Carroll, J. S. (2011).
Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1.
The EMBO Journal, 30(15):3019–27.

Rocha, G., Zhao, P., and Yu, B. (2008).
A path following algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE).
Technical report, Statistics Department, UC Berkeley, Berkeley, CA.

Sauter, E. R., Lininger, J., Magklara, A., Hewett, J. E., and Diamandis, E. P. (2004).
Association of kallikrein expression in nipple aspirate fluid with breast cancer risk.
International Journal of Cancer, 108(4):588–91.

Shimo, A., Nishidate, T., Ohta, T., Fukuda, M., Nakamura, Y., and Katagiri, T. (2007).
Elevated expression of protein regulator of cytokinesis 1, involved in the growth of breast cancer cells.
Cancer Science, 98(2):174–81.

Shimo, A., Tanikawa, C., Nishidate, T., Lin, M.-L., Matsuda, K., Park, J.-H., Ueki, T., Ohta, T., Hirata, K., Fukuda, M., Nakamura, Y., and Katagiri, T. (2008).
Involvement of kinesin family member 2C/mitotic centromere-associated kinesin overexpression in mammary carcinogenesis.
Cancer Science, 99(1):62–70.

Sizemore, S. T. and Keri, R. A. (2012).
The Forkhead Box Transcription Factor FOXC1 Promotes Breast Cancer Invasion by Inducing Matrix Metalloprotease 7 (MMP7) Expression.
The Journal of Biological Chemistry, 287(29):24631–40.

References VI

Tkocz, D., Crawford, N. T., Buckley, N. E., Berry, F. B., Kennedy, R. D., Gorski, J. J., Harkin, D. P., and Mullan, P. B. (2012).
BRCA1 and GATA3 corepress FOXC1 to inhibit the pathogenesis of basal-like breast cancers.
Oncogene, 31(32):3667–3678.

Wang, J., Ray, P. S., Sim, M.-S., Zhou, X. Z., Lu, K. P., Lee, A. V., Lin, X., Bagaria, S. P., Giuliano, A. E., and Cui, X. (2012).
FOXC1 regulates the functions of human basal-like breast cancer cells by activating NF-κB signaling.
Oncogene.

Yan, W., Cao, Q. J., Arenas, R. B., Bentley, B., and Shao, R. (2010).
GATA3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition.
The Journal of Biological Chemistry, 285(18):14042–14051.

Yang, Q., Nakamura, M., Nakamura, Y., Yoshimura, G., Suzuma, T., Umemura, T., Tamaki, T., Mori, I., Sakurai, T., and Kakudo, K. (2002).
Correlation of prostate-specific antigen promoter polymorphisms with clinicopathological characteristics in breast cancer.
Anticancer Research, 22(3):1825–8.

Zheng, Y., Huo, D., Zhang, J., Yoshimatsu, T. F., Niu, Q., and Olopade, O. I. (2012).
Microsatellites in the Estrogen Receptor (ESR1, ESR2) and Androgen Receptor (AR) Genes and Breast Cancer Risk in African American and Nigerian Women.
PLoS ONE, 7(7):e40494.

Simulation: Pseudo-likelihood methods

Datasets:

- True Ω has 2.4% non-zero elements (placed at random)

$$\mathbf{Y} \sim \mathcal{N}_{100}(0, \Omega^{-1})$$

- Generate 100 independent datasets, $p = 100$, $n = 200$
- Grid of 50 penalty parameter (λ) values

Model selection performance metrics:

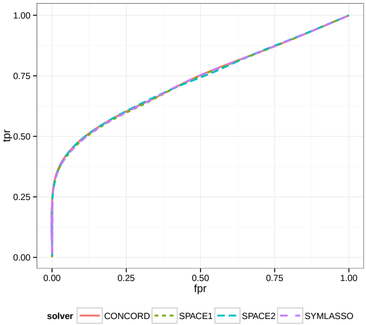
- Measures performance of **zero vs. non-zero structure** recovery
- False Positive Rate (FPR) vs. True Positive Rate (TPR):

$$FPR = \frac{FP}{FP + TN} \quad \text{and} \quad TPR = \frac{TP}{TP + FN}$$

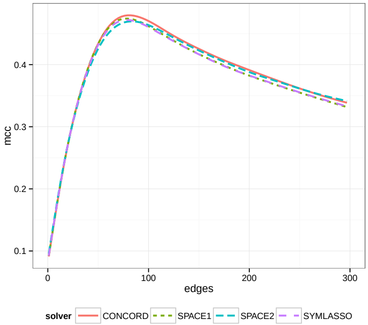
- # of non-zeros vs. Matthew's correlation coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Simulation: Pseudo-likelihood methods



(a) FPR vs. TPR



(b) # of non-zeros in $\hat{\Omega}$ vs. MCC

CONCORD has competitive model selection performance

Insights for CONCORD: part 1

$$\mathcal{L}_{\text{con}}(\Omega_D, \Omega^2) = \frac{n}{2}\left[-\log\det\Omega_D + \text{tr}(\mathbf{S}\Omega^2)\right]$$

- $\log|\Omega| \longrightarrow \log|\Omega_D|$
- $\text{tr}(\mathbf{S}\Omega) \longrightarrow \text{tr}(\Omega\mathbf{S}\Omega) = \text{tr}(\mathbf{S}\Omega^2)$
- Modify the log determinant term to balance

$$\mathcal{L}_{\text{uni}}(\Omega_D^2, \Omega^2) = \frac{n}{2}\left(-\log\det\Omega_D^2 + \text{tr}(\mathbf{S}\Omega^2)\right)$$

- Penalized pseudo-likelihood of CONCORD

$$Q_{\text{con}}(\Omega) := \mathcal{L}_{\text{uni}}(\Omega_D^2, \Omega^2) + \lambda\sum_{i<j}|\omega_{ij}|$$

- Modification gives better parameter estimates

Insights for CONCORD: part 2

Generated Gaussian dataset with following Ω^* $(n = 1000)$.

$$\Omega^* = \begin{pmatrix} 1.0 & 0.3 & 0.0 \\ 0.3 & 1.0 & 0.3 \\ 0.0 & 0.3 & 1.0 \end{pmatrix}$$

For $\lambda = 0$,

$$\Omega_{\text{uncorrected}} = \begin{pmatrix} 0.675 & 0.089 & -0.015 \\ 0.089 & 0.658 & 0.117 \\ -0.015 & 0.117 & 0.668 \end{pmatrix}$$

$$\Omega_{\text{con}} = \begin{pmatrix} 0.974 & 0.257 & 0.007 \\ 0.257 & 0.983 & 0.344 \\ 0.007 & 0.344 & 0.978 \end{pmatrix}$$

Modified likelihood gives better parameter estimates!