Sparse Gaussian Graphical Model Estimation using ℓ_1 Regularization

Ming Yuan

School of Industrial and Systems Engineering

Georgia Institute of Technology

myuan@isye.gatech.edu

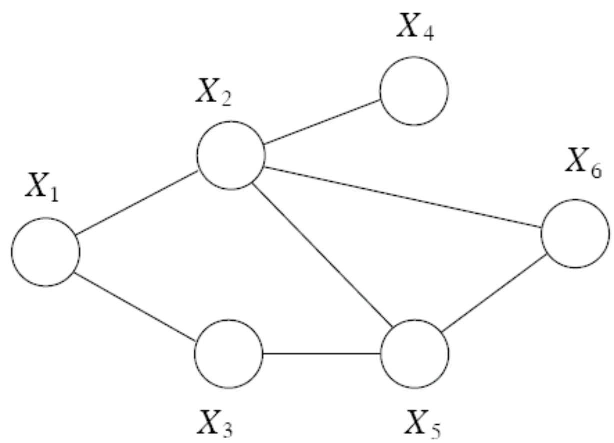
http://www.isye.gatech.edu/~myuan

OUTLINE

- Introduction
- Methodology
- Asymptotic Properties
- Computation
- Numerical Examples

BACKGROUND

UNDIRECTED GRAPHICAL MODEL



- $X_{\mathcal{V}}$ is represented by an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$
  - ◇ $\mathcal{V} = \{1, 2, 3, 4, 5, 6\}$ contains vertices corresponding to the random variables
  - ◇ the edges $\mathcal{E} = \{(1, 2), (1, 3), \ldots, (5, 6)\}$
- Factorization of probability distribution

$$p(\mathbf{x}_{\mathcal{V}}) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{25}(x_2, x_5)\psi_{26}(x_2, x_6)\psi_{35}(x_3, x_5)\psi_{56}(x_5, x_6)$$

- Conditional independence

$$X_2 \perp X_3 \Big| X_1, X_4, X_5, X_6$$

Gaussian Graphical Model

$$X = (X^{(1)}, \ldots, X^{(p)}) \sim \mathcal{N}_p(\mu, \Sigma)$$

- Concentration matrix $C = \Sigma^{-1}$

$$\begin{aligned} p(\mathbf{x}_{\mathcal{V}}) &= (2\pi)^{-p/2} |C|^{1/2} \exp\left\{ - \sum_{(i,j):c_{ij}\neq 0} c_{ij}(x_i - \mu_i)(x_j - \mu_j)/2 \right\} \\ &= \underbrace{(2\pi)^{-p/2} |C|^{1/2}}_{1/Z} \prod_{(i,j):c_{ij}\neq 0} \underbrace{\exp\{-c_{ij}(x_i - \mu_i)(x_j - \mu_j)/2\}}_{\psi_{ij}(x_i,x_j)} \end{aligned}$$

- Graphical Model – $\mathcal{E} = \{(i,j) : c_{ij} \neq 0\}$:

$$c_{ij} = 0 \Longrightarrow X_i \perp X_j \big| X_{-\{i,j\}}$$

- Goal
  - ◇ Covariance Selection – estimate the graphical structure $\mathcal{E}$
  - ◇ Covariance Estimation – estimate Σ – PCA, LDA and etc.

Time Series

- Markov Property

$$C = \begin{bmatrix} c_{11} & c_{12} & 0 & \cdots & \cdots & 0 \\ c_{21} & c_{22} & c_{23} & \ddots & \ddots & \vdots \\ 0 & c_{32} & c_{33} & c_{34} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & c_{p-1,p-2} & c_{p-1,p-1} & c_{p-1,p} \\ 0 & \cdots & \cdots & 0 & c_{p,p-1} & c_{pp} \end{bmatrix}$$

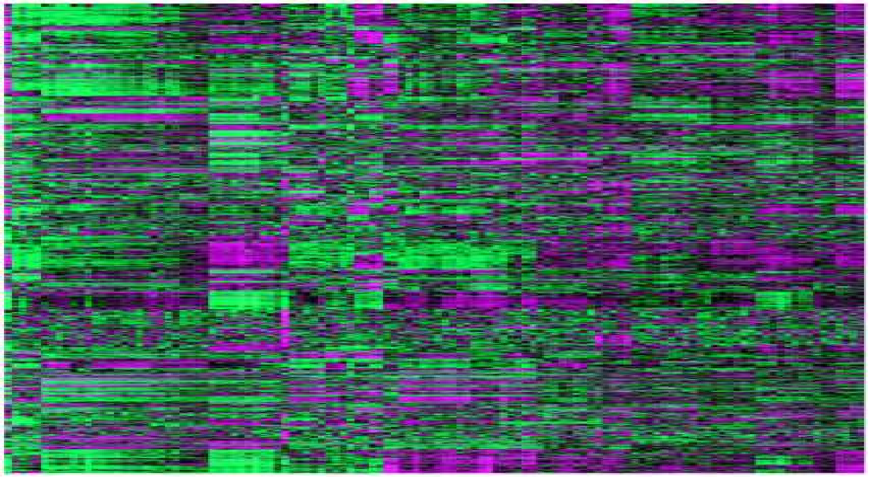
IMAGES

- Markov random fields

$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
$X^{(4)}$	$X^{(5)}$	$X^{(6)}$
$X^{(7)}$	$X^{(8)}$	$X^{(9)}$

GENE NETWORK AND PATHWAY

(a) *Arabidopsis*



(b) Gene Expression of 835 Genes

Isoprenoids ⟵ Five-carbon intermediates (IPP, DMAPP) ⟵ MEP Pathway, MVA Pathway

How to do it

- Classical approach (Lauritzen, 1996; Whittaker, 1990)
  - ◇ Greedy stepwise forward-selection or backward-deletion
  - ◇ the edge selection or deletion through hypothesis testing
  - ◇ computational complexity
  - ◇ does not correctly adjust for the multiple comparisons
- Recent Advances
  - ◇ Drton and Perlman (2004)
    - ☞ produces conservative simultaneous $1 - \alpha$ confidence intervals
    - ☞ uses these confidence intervals to do model selection in a single step
  - ◇ Meinshausen and Bühlmann (2006) – neighbourhood selection
    - ☞ Lasso for each node to identify its neighbours
    - ☞ combine the results to learn the structure of a Gaussian concentration graph model
  - ◇ Bayesian approaches: Liechty, Liechty and Müller (2004); Dobra and West (2005)

ℓ_1 Regularization

CONSTRAINED MAXIMUM LIKELIHOOD ESTIMATE

$$C \equiv \Sigma^{-1} \Rightarrow e_{ij} = I(c_{ij} \neq 0)$$

- To find a sparse graph

$$\min_{C \succ 0} \left[-\ln |C| + \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)' C (X_i - \mu) \right] \quad \text{subject to} \quad \sum_{i \neq j} |e_{ij}| \leq t$$

- Lasso-type Estimate – Graphical Lasso

$$\min_{C \succ 0} \left[-\ln |C| + \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)' C (X_i - \mu) \right] \quad \text{subject to} \quad \sum_{i \neq j} |c_{ij}| \leq t$$

- Graphical Garrote – $\tilde{C}$ a preliminary estimate $\Rightarrow c_{ij} = d_{ij} \tilde{c}_{ij}$

$$\min_{C \succ 0} \left[-\ln |C| + \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)' C (X_i - \mu) \right] \quad \text{subject to} \quad \sum_{i \neq j} d_{ij} \leq t, d_{ij} \geq 0,$$

Relationship to Semi-definite Programming

MaxDet Problem

$$\begin{aligned} \min_{x \in R^m} \quad & b'x - \ln |G(x)| \\ \text{subject to} \quad & G(x) \text{ is positive definite} \\ & F(x) \text{ is positive semi} - \text{definite} \end{aligned}$$

where $b \in R^m$ and the functions $G : R^m \to R^{l \times l}$ and $F : R^m \to R^{l \times l}$ are affine:

$$\begin{aligned} G(x) &= G_0 + x_1 G_1 + \ldots + x_m G_m, \\ F(x) &= F_0 + x_1 F_1 + \ldots + x_m F_m \end{aligned}$$

where F_i and G_i are symmetric matrices such that

- $F_i, i = 1, \ldots, m$ are linearly independent
- $G_i, i = 1, \ldots, m$ are linearly independent

PROPERTIES

Lagrange Form

$\bar{A}$ – MLE of the sample covariance matrix

- Lasso-type estimate

$$\min \left[- \ln |C| + \operatorname{trace}\left(C\bar{A} \right) + \lambda \|C\|_{\ell_1} \right]$$

where $\|C\|_{\ell_1} = \sum_{i \neq j} |c_{ij}|$.

- Garrote-type estimate

$$\min \left[- \ln |C| + \operatorname{trace}\left(C\bar{A} \right) + \lambda \sum_{i \neq j} \frac{c_{ij}}{\tilde{c}_{ij}} \right]$$

subject to $c_{ij}/\tilde{c}_{ij} \geq 0$.

BIVARIATE NORMAL

$$\widehat{C}_0 = \bar{A}^{-1} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

THEOREM *In the case of the bivariate normal, the Lasso-type estimate is*

$$\widehat{c}_{12} = \left(\frac{(1-r^2)\{|r| - \lambda(1-r^2)\}}{1 - \{|r| - \lambda(1-r^2)\}^2} \right)_+ \text{sign}(r),$$

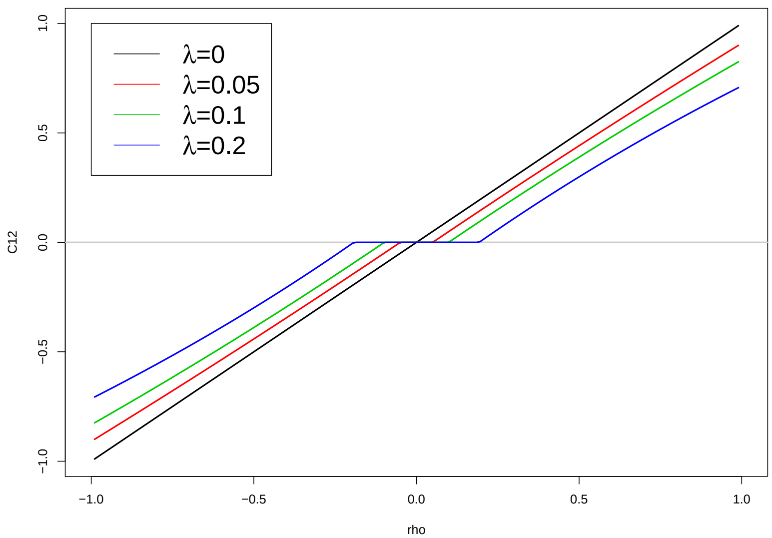
the Garrote-type estimate is

$$\widehat{c}_{12} = \left(\frac{(1-r^2)\{r^2 - \lambda(1-r^2)\}}{|r| - \{r^2 - \lambda(1-r^2)\}^2/|r|} \right)_+ \text{sign}(r)$$

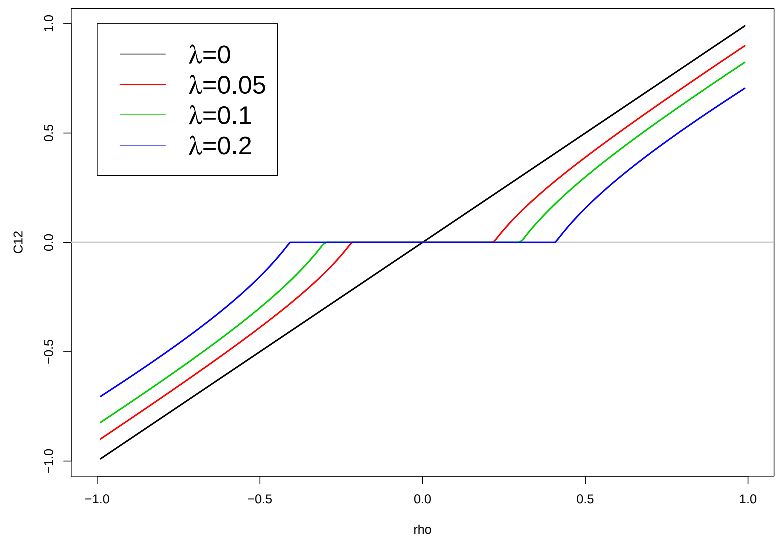
and

$$\widehat{c}_{11} = \widehat{c}_{22} = \frac{1}{2}\left[(1-r^2) + \sqrt{\{(1-r^2)^2 + 4\widehat{c}_{12}^2\}}\right].$$

BIVARIATE NORMAL



(c) Lasso Type Estimate



(d) Garrote Type Estimate

- both estimates of c_{12} can be shrunken to exact zero
- both estimates are continuous in the data
- The garrote type estimate penalizes large r's less heavily

ASYMPTOTICS – GRAPHICAL GARROTE

THEOREM *Denote by* $\widehat{C}$ *the garrote-type estiamte with initial estimator* $\tilde{C} = \bar{A}^{-1}$*. If* $n\lambda \to \infty$ *and* $\sqrt{n}\lambda \to 0$ *as* $n \to \infty$*, then* $\Pr(\widehat{c}_{ij} = 0) \to 1$ *if* $c_{ij} = 0$*, and other elements of* $\widehat{C}$ *have the same limiting distribution as the maximum likelihood estimator on the true graph structure.*

- the garrote type estimate enjoys the so-called oracle property
- needs a good preliminary estimate

Asymptotics – Graphical Lasso

THEOREM *If* $\sqrt{n}\lambda \to \lambda_0 \geq 0$ *as* $n \to \infty$, *the lasso-type estimator is such that*

$$\sqrt{n}\left(\widehat{C} - C\right) \to \arg\min_{U=U'}(V),$$

in distribution, where

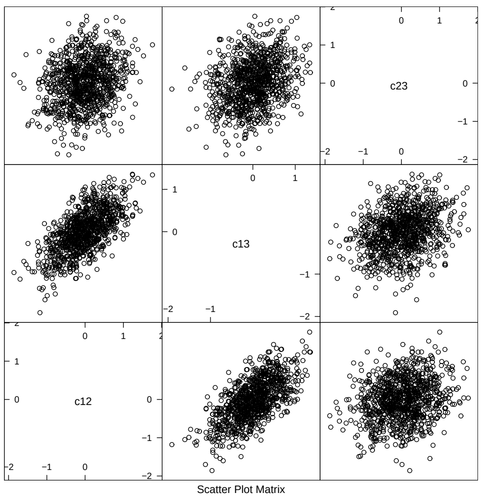
$$\begin{aligned} V(U) &= \text{trace}\,(U\Sigma U\Sigma) + \text{trace}(UW) + \\ &\quad \lambda_0 \sum_{i\neq j}\{u_{ij}\text{sign}(c_{ij})I(c_{ij}\neq 0) + |u_{ij}|I(c_{ij}=0)\} \end{aligned}$$

in which W *is a random symmetric* $p \times p$ *matrix such that* $\text{vec}(W) \sim \mathcal{N}(0,\Lambda)$, *and* Λ *is such that*

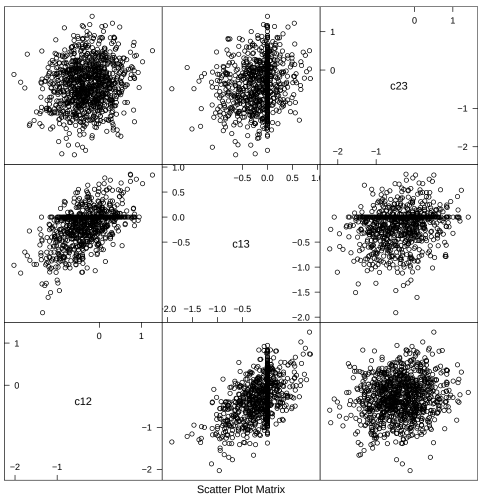
$$\text{cov}(w_{ij}, w_{i'j'}) = \text{cov}(X^{(i)}X^{(j)}, X^{(i')}X^{(j')}).$$

EXAMPLE – TRIVARIATE NORMAL

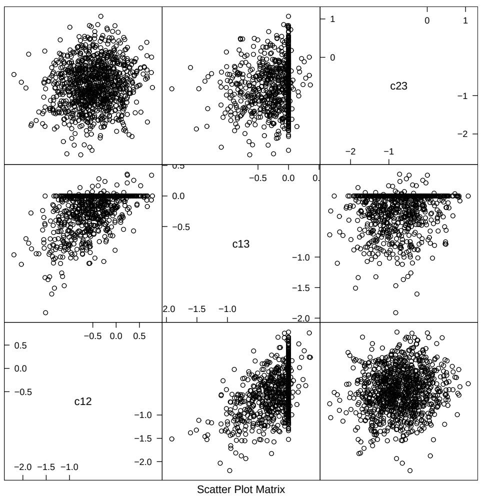
$$C = \begin{pmatrix} 1 & \frac{1}{3} & 0 \\ \frac{1}{3} & 1 & \frac{2}{3} \\ 0 & \frac{2}{3} & 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1.25 & -0.75 & 0.5 \\ -0.75 & 2.25 & -1.5 \\ 0.5 & -1.5 & 2 \end{pmatrix}.$$



(a) $\lambda_0 = 0$



(b) $\lambda_0 = 0.5$



(c) $\lambda_0 = 1$

COMPUTATION

Coordinate Descent

- If $C_{-i,-i} \succ 0$, then $C \succ 0$ iff

$$c_{ii} - C_{i,-i} C_{-i,-i}^{-1} C_{-i,i} > 0$$

- Given $C_{-i,-i} \succ 0$, then $C \succ 0$,

$$\min_{C_{i,\cdot} = C'_{\cdot,i}} \left[-\ln |C| + \text{trace}\left(C\bar{A}\right) + \lambda \sum_{i \neq j} |c_{ij}| \right]$$

subject to

$$c_{ii} - C_{i,-i} C_{-i,-i}^{-1} C_{-i,i} > 0$$

BACKFITTING

- In terms of $C_{\cdot,i}$,

$$-\ln\left(c_{ii} - C'_{-i,i}C^{-1}_{-i,-i}C_{-i,i}\right) + c_{ii}\bar{A}_{ii} + 2\bar{A}_{i,-i}C_{-i,i} + 2\lambda\sum_{j\neq i}|c_{ij}|$$

- First order condition
  - ◇ Diagonal element

$$c_{ii} = \frac{1}{\bar{a}_{ii}} + C'_{-i,i}C^{-1}_{-i,-i}C_{-i,i}$$

  - ◇ Off-diagonal elements

$$\frac{1}{2}C'_{-i,i}\left(\bar{a}_{ii}C^{-1}_{-i,-i}\right)C_{-i,i} + A_{i,-i}C_{-i,i} + \lambda\sum_{j\neq i}|c_{ij}|$$

Graphical Lasso Algorithm

Input: $\bar{A}$, $\lambda \geq 0$ and an initial value for C

Output: Update for the ith row and column

Repeat

for $i = 1$ to p

update $C_{-i,i}$, or equivalently $C_{i,-i}$ by solving

$$\min_{C_{-i,i}=C'_{i,-i}} \frac{1}{2} C'_{-i,i} \left(\bar{a}_{ii} C^{-1}_{-i,-i}\right) C_{-i,i} + \bar{A}_{i,-i} C_{-i,i} + \lambda \sum_{j \neq i} |c_{ij}|$$

update $c_{ii} = 1/\bar{a}_{ii} + C'_{-i,i} C^{-1}_{-i,-i} C_{-i,i}$

end

Until certain convergence criterion is met

Graphical Garrote Algorithm

Input: $\bar{A}$, $\lambda \geq 0$ and an initial value for C

Output: Update for the ith row and column

Repeat

for $i = 1$ to p

update $C_{-i,i}$, or equivalently $C_{i,-i}$ by solving

$$\min_{C_{-i,i}=C'_{i,-i}} \frac{1}{2} C'_{-i,i} \left(\bar{a}_{ii} C^{-1}_{-i,-i}\right) C_{-i,i} + A_{i,-i} C_{-i,i} + \lambda \sum_{j \neq i} \frac{c_{ij}}{\tilde{c}_{ij}} \quad \text{subject to } c_{ij}/\tilde{c}_{ij} \geq 0$$

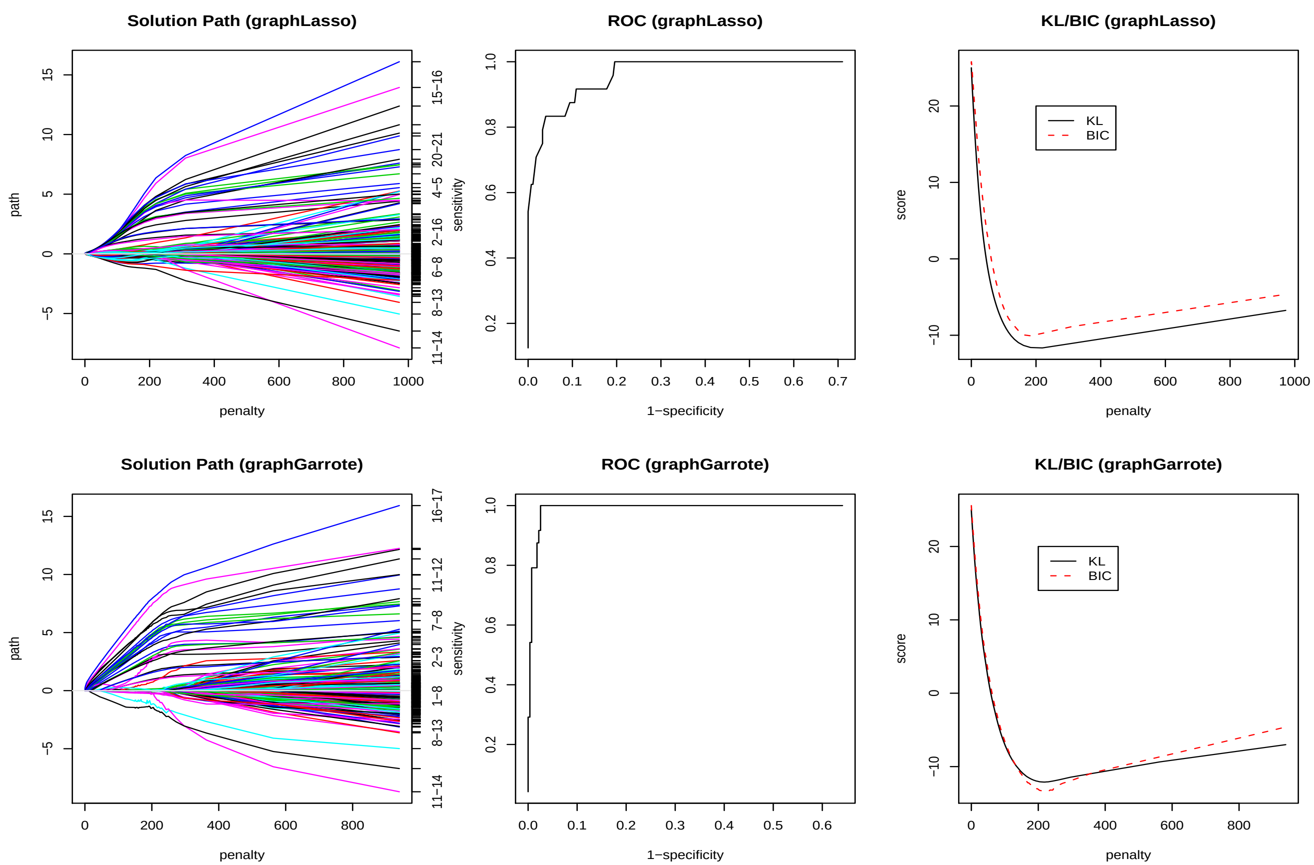
update $c_{ii} = 1/\bar{a}_{ii} + C'_{-i,i} C^{-1}_{-i,-i} C_{-i,i}$

end

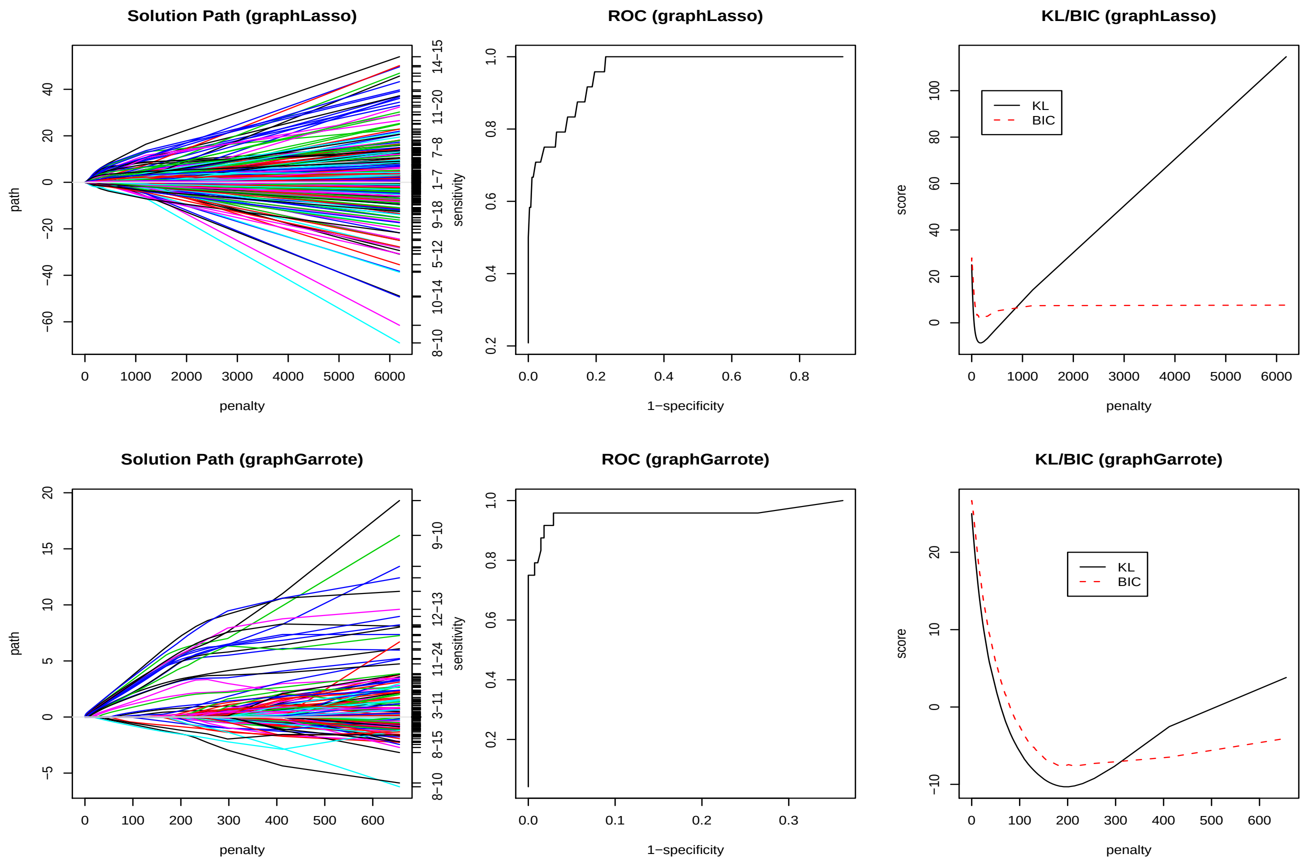
Until certain convergence criterion is met

NUMERICAL EXAMPLES

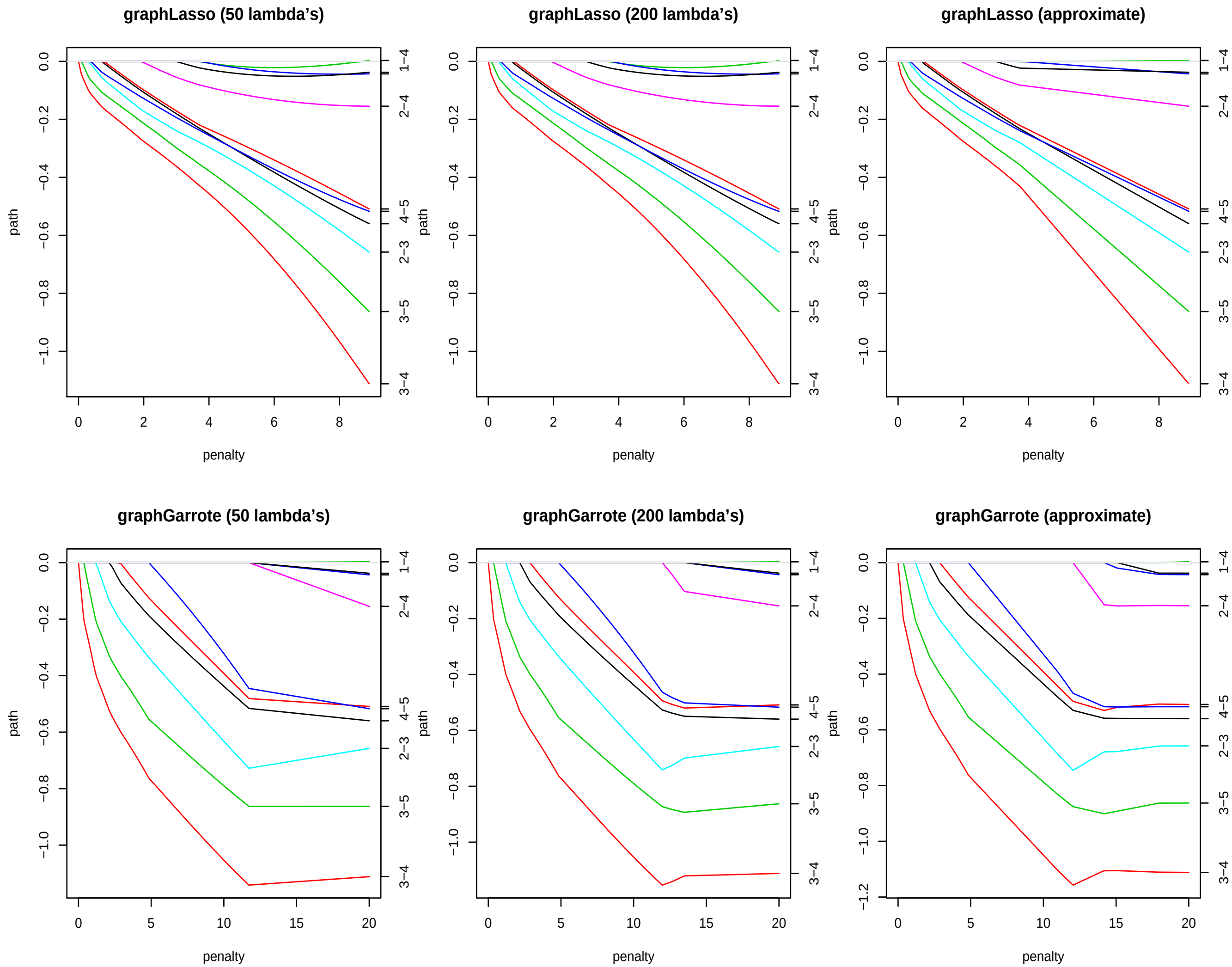
AR(1) Model – $p = 25, n = 100$



AR(1) MODEL – $n = p = 25$

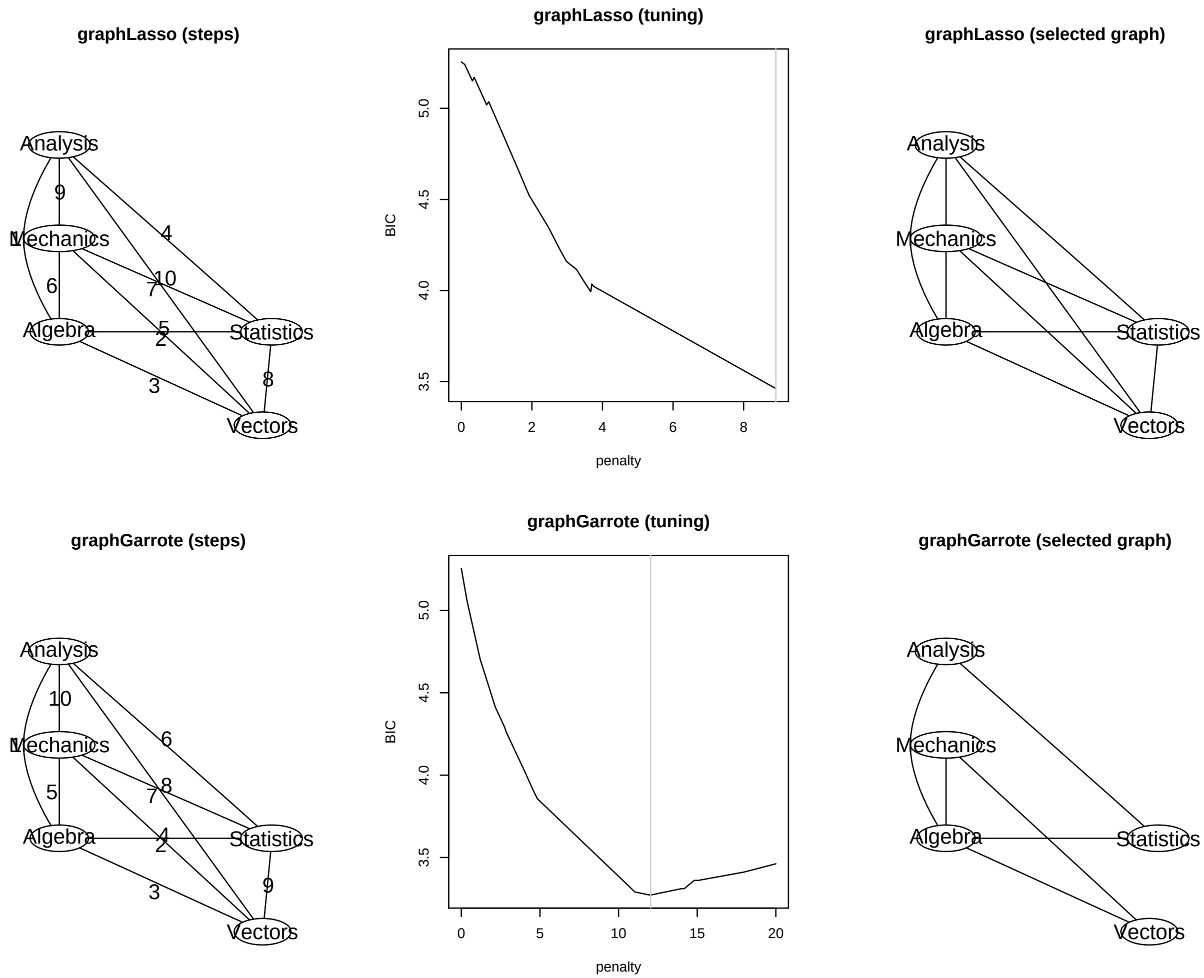


MATHMARKS DATA

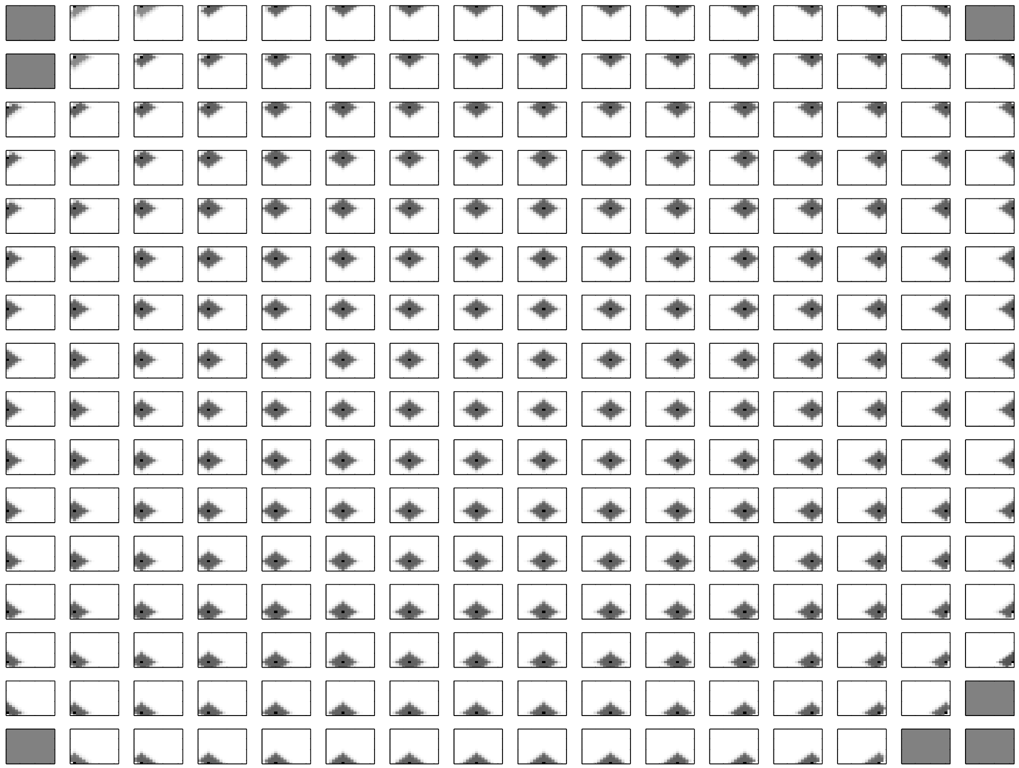


Order	graphLasso	graphGarrote
Step 1	Algebra – Analysis	Algebra – Analysis
Step 2	Algebra – Statistics	Algebra – Statistics
Step 3	Vector – Algebra	Vector – Algebra
Step 4	Analysis – Statistics	Mechanics – Vector
Step 5	Mechanics – Vector	Mechanics – Algebra
Step 6	Mechanics – Algebra	Analysis – Statistics
Step 7	Vector – Analysis	Vector – Analysis
Step 8	Vector – Statistics	Mechanics – Statistics
Step 9	Mechanics – Analysis	Vector – Statistics
Step 10	Mechanics – Statistics	Mechanics – Analysis

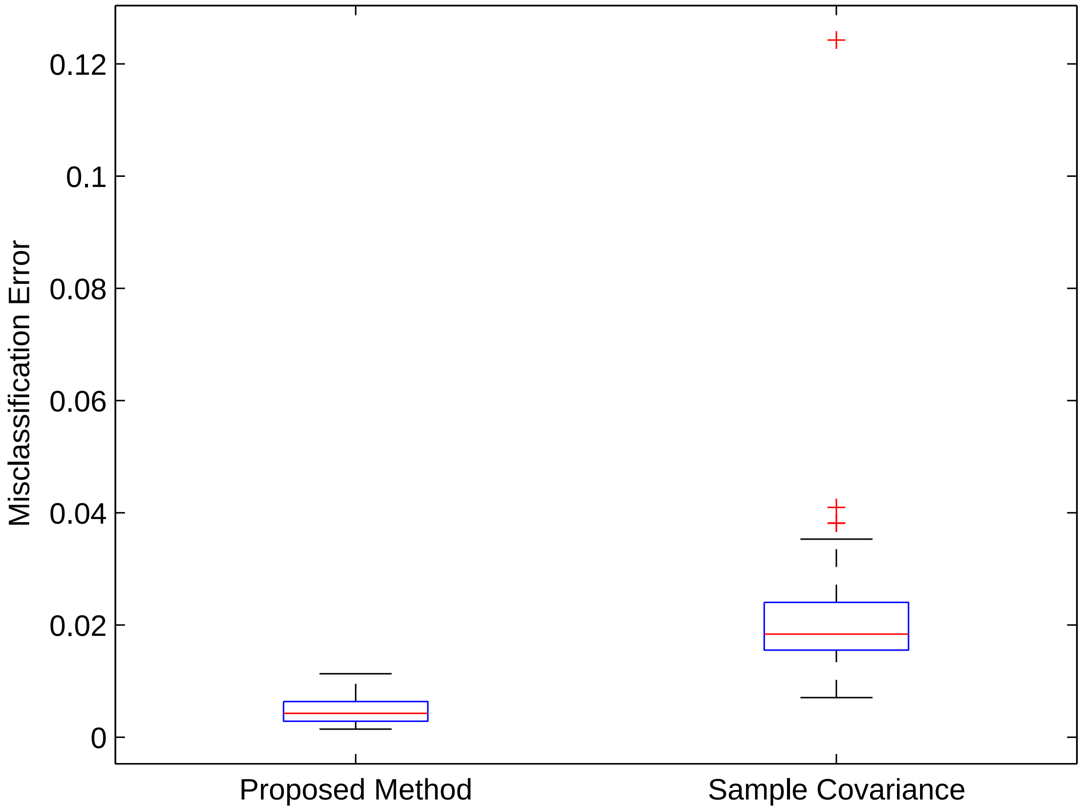
MATH MARKS DATA



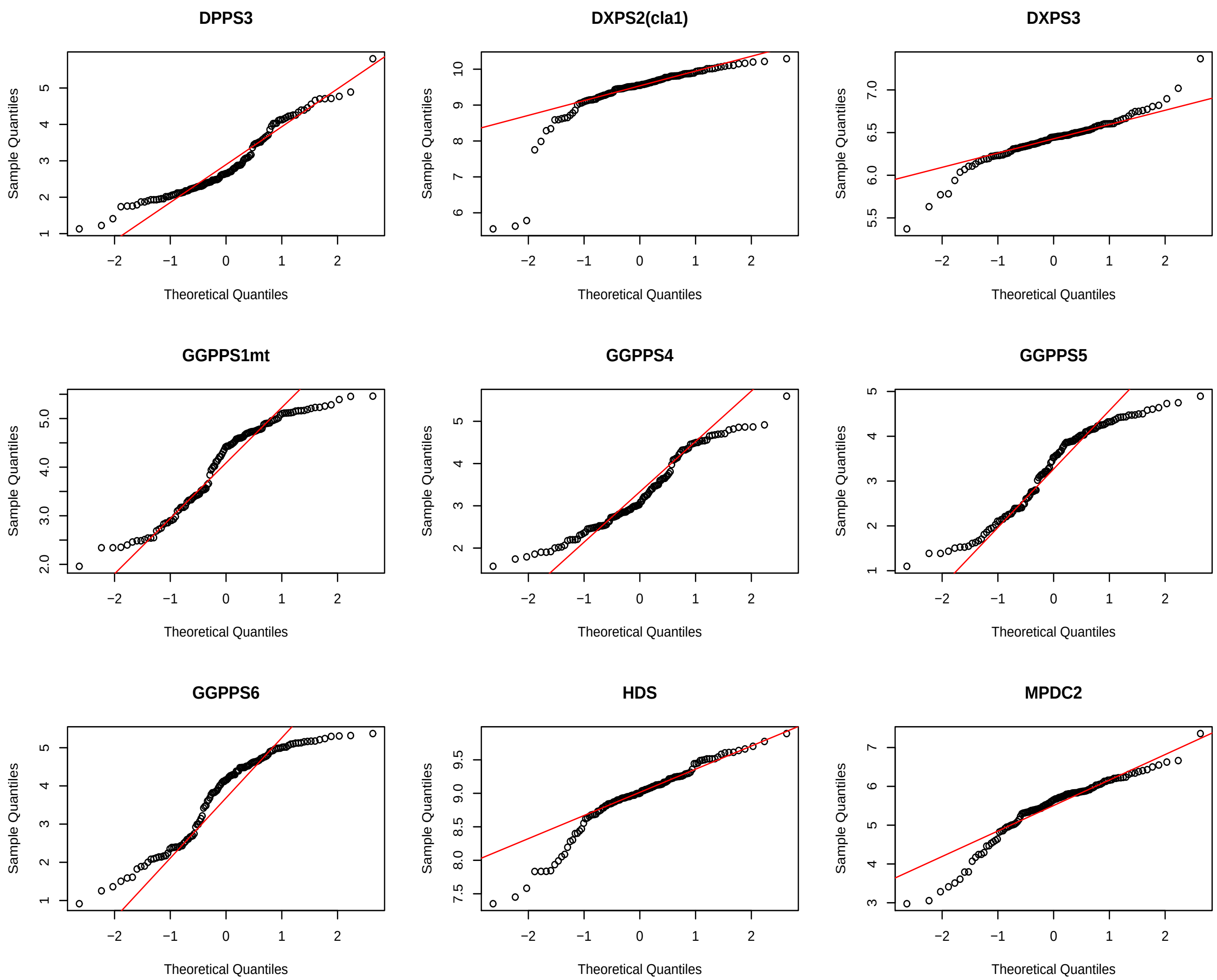
Handwritten Digits Classification



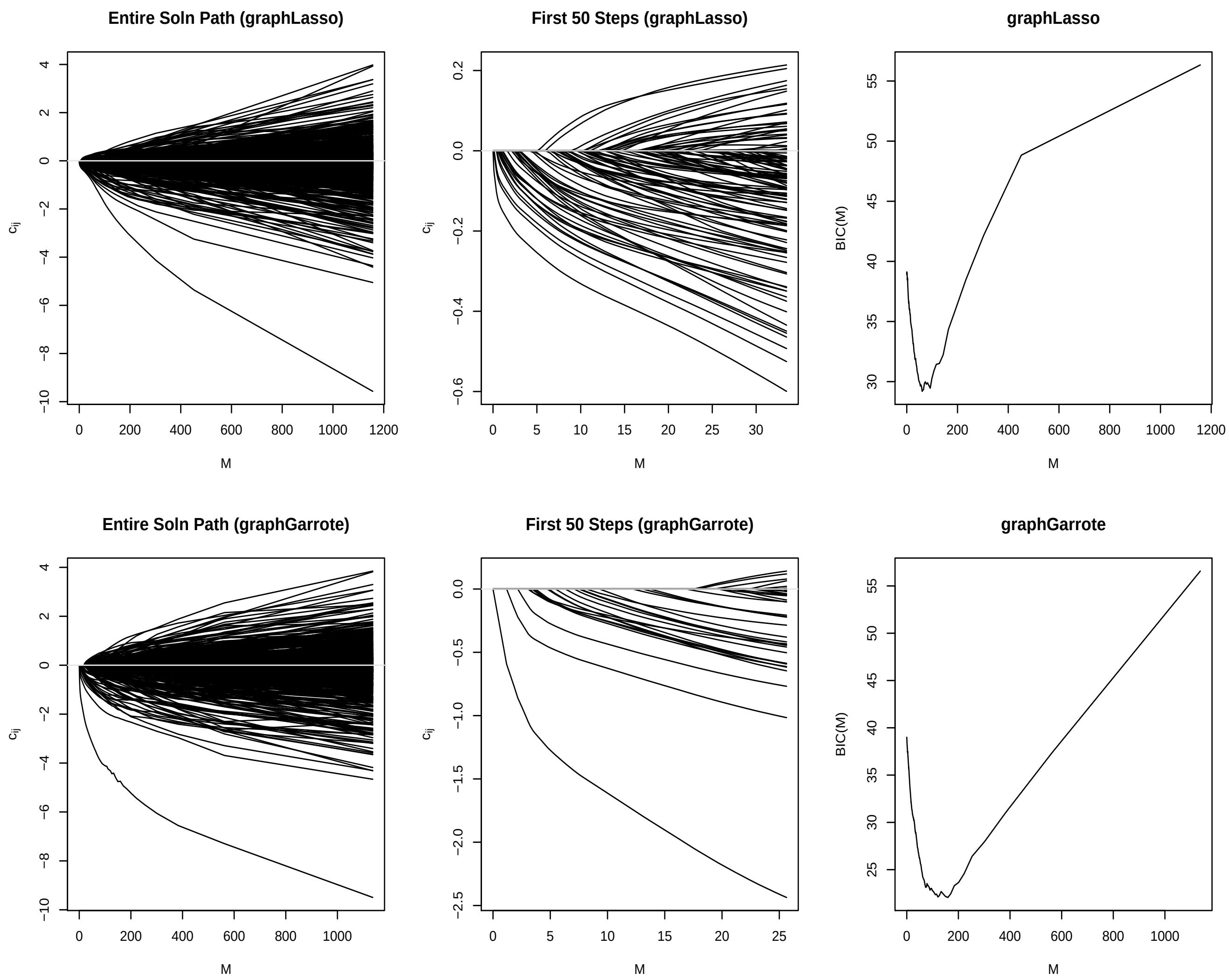
Handwritten Digits Classification



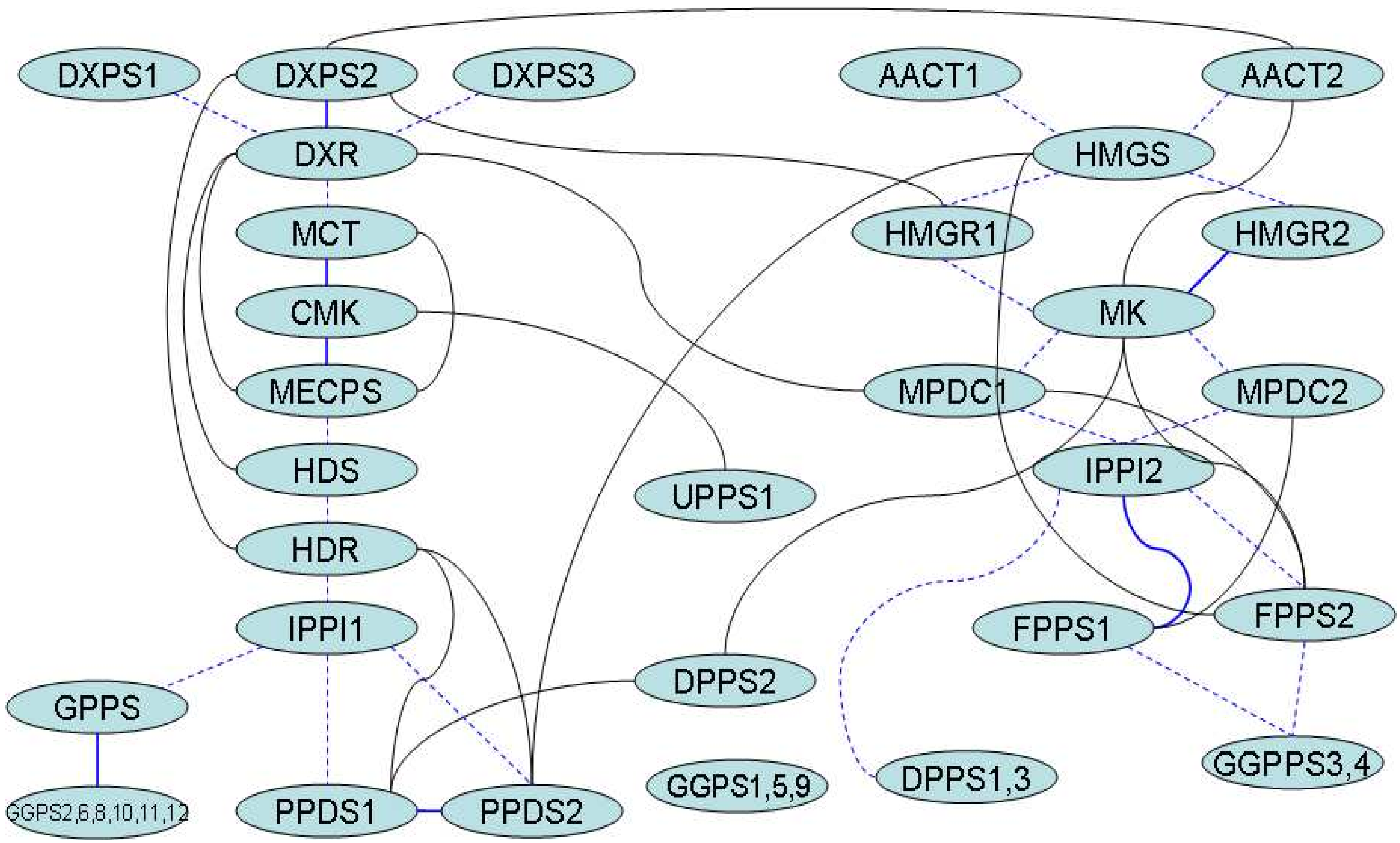
THALIANA DATA



Thaliana Data



Thaliana Data



CONCLUSION

- We have introduced ℓ_1 regularized likelihood methods for the problem of covariance selection and estimation
- The implementation of our method takes advantage of recent advances in convex optimization
- Numerical experiments demonstrate the merits of the new method

Georgia Institute of Technology