Disc-GLasso: Discriminative Graph Learning with Sparsity Regularization

Jiun-Yu Kao[1], **Dong Tian**[2], **Hassan Mansour**[2], **Antonio Ortega**[1] **and Anthony Vetro**[2]

[1] University of Southern California, Los Angeles, CA

[2] Mitsubishi Electric Research Labs (MERL), Cambridge, MA

Motivation

Graph structures are natural to use

- Construct optimal graph not trivial
  - benefit subsequent data analysis & could learn from data
- Prior art learn graphs w.r.t.
  - Representability: benefits energy compaction
  - Sparsity: benefits interpretation

Objectives

- Multi-class classification among data samples (graph signals)
- Develop approach to learn a set of graphs that benefits classification

Key Contributions

- First propose multi-graph learning that promotes discrimination
- Develop efficient algorithm to learn discriminative class-specific graphs

Problem Formulation

Notations

- random graph signals in i-th class: $\mathbf{x}^{(i)} \in \mathbb{R}^n$
- S classes in total
- For i-th class, N_i i.i.d. realizations/samples: $\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_{N_i}^{(i)}$
- $\mathbf{K}_i$: empirical cov. matrix of samples in i-th class

Define multi-category graph learning problem:

Goal: learn graph structure of each category, $\mathcal{G}_1, \ldots, \mathcal{G}_S$

- $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}_i, \mathbf{Q}_i)$, n-vertex set $\mathcal{V}$, edge set $\mathcal{E}_i$
- Symmetric matrix representation $\mathbf{Q}_i$, where $\forall a \neq b$, $(a,b) \in \mathcal{E}_i \iff \mathbf{Q}_{i,ab} \neq 0$
- $\mathbf{Q}_i$ is only required to be positive semi-definite

Baseline Approach

Apply graphical lasso indep. to each class:

- Solve ℓ_1-penalized Gaussian ML estimation problem separately for the graph of each category i,

$$\min -\log\det(\mathbf{Q}_i) + \mathrm{tr}(\mathbf{K}_i\mathbf{Q}_i) + \rho\|\mathbf{Q}_i\|_1$$

  - $\mathrm{tr}(\mathbf{K}_i\mathbf{Q}_i) = \frac{1}{N_i}\sum_{k=1}^{N_i} \mathbf{x}_k^{(i)^T}\mathbf{Q}_i\mathbf{x}_k^{(i)}$
  - Minimize above term $\Rightarrow$ promote smoothness of data samples in i-th class on i-th graph
- Only favor energy compaction and sparsity within each class
- Discrimination between classes not guaranteed

Disc-GLasso Algorithm

For each $\mathbf{W}_i$, given the empirical covariance matrices $\mathbf{K}_1, \cdots, \mathbf{K}_S$.

1. Search for the minimum ratio r such that $\mathbf{K}_i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j \succeq 0, \forall i$.
2. Initialize with $\mathbf{W}_i = \mathbf{K}_i + \rho\mathbf{I} - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j$. The diagonal of $\mathbf{W}_i$ will remain unchanged in what follows.
3. Perform following steps until convergence reached:
   (a) Rearrange the rows/columns so that the target column is last.
   (b) Solve the lasso problem for $\hat{\beta}$:

$$\underset{\beta}{\mathrm{argmin}}\frac{1}{2}\left\|\mathbf{W}_{11}^{i\,1/2}\beta - \mathbf{W}_{11}^{i\,-1/2}\mathbf{k}_{12}^{i} + \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{W}_{11}^{i\,-1/2}\mathbf{k}_{12}^{j}\right\|^2 + \rho\|\beta\|_1$$

   (c) Fill in the corresponding row and column of $\mathbf{W}_i$ using $\mathbf{w}_{12}^i = \mathbf{W}_{11}^i\hat{\beta}$.

Proposed Solution

- Key idea: jointly learn the graphs for all classes by promoting
  - Smoothness of data in i-th class on i-th graph
  - Non-smoothness on learned graphs corresponding to the other classes
- To achieve above properties, we need to minimize the following,

$$\sum_{i=1}^{S}\sum_{k=1}^{N_i}\frac{1}{N_i}\left[\mathbf{x}_k^{(i)^T}\mathbf{Q}_i\mathbf{x}_k^{(i)} - \frac{1}{S-1}\sum_{j\neq i}^{S}\mathbf{x}_k^{(i)^T}\mathbf{Q}_j\mathbf{x}_k^{(i)}\right]$$

- Finally, we propose to solve the optimization problem below.

$$\min_{\mathbf{Q}_i\succeq 0} -\log\det(\mathbf{Q}_i) + \mathrm{tr}(\mathbf{K}_i\mathbf{Q}_i) - \frac{\mu_i}{S-1}\sum_{j\neq i}^{S}\mathrm{tr}(\mathbf{K}_j\mathbf{Q}_i) + \rho_i\|\mathbf{Q}_i\|_1, \quad (1)$$

  for each $\mathbf{Q}_i$, given $\mathbf{K}_1, \ldots, \mathbf{K}_S$.
- A block coordinate descent based algorithm is developed to solve (1).

Parameter Choice: Let $r = \frac{S-1}{\mu}$,

- Smaller $r \Rightarrow$ better discrimination
- Need to ensure $\mathbf{K}_i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j \succeq 0, \forall i$ to guarantee that after each updating step t, $\mathbf{W}_i^{(t)} \succ 0, \forall t$.

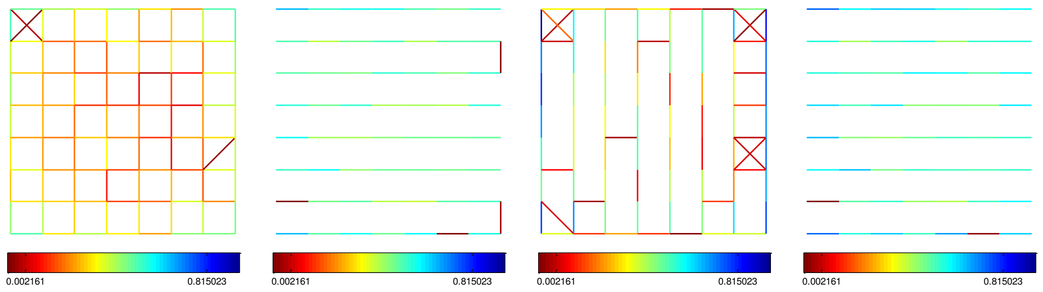
In proposed algorithm, we search for the minimum r s.t. $\mathbf{K}_i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j \succeq 0, \forall i$, via line search through predefined set of values.

- the only additional time cost of our algorithm compared to GLasso

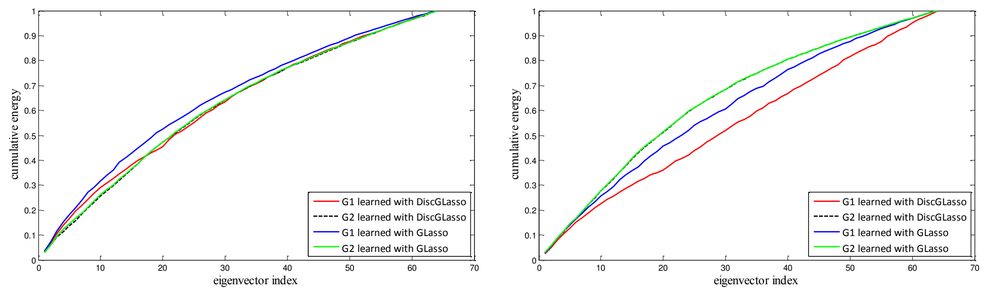
Experiments

Synthetic data for binary classification

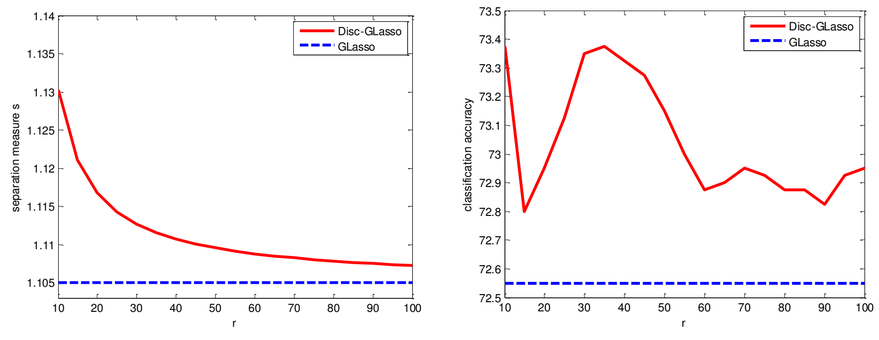
- **From left to right:** $\mathcal{G}_1$, $\mathcal{G}_2$ learned with GLasso, and $\mathcal{G}_1$, $\mathcal{G}_2$ learned with Disc-GLasso.



- Cumulative spectrum energy of test signals in class 1 **(left)** and class 2 **(right)** on the learned graphs.



- Effect of r on the separation measure $s = \frac{\mathrm{tr}(\mathbf{K}_1\mathbf{Q}_2)+\mathrm{tr}(\mathbf{K}_2\mathbf{Q}_1)}{\mathrm{tr}(\mathbf{K}_1\mathbf{Q}_1)+\mathrm{tr}(\mathbf{K}_2\mathbf{Q}_2)}$ **(left)** and classification accuracy **(right)**.



References

1. J. Friedman et.al, "Sparse inverse covariance estimation with the graphical lasso," Biostatistics, 2008.
2. R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," Elect. journal of stat., 2012.