DISC-GLASSO: DISCRIMINATIVE GRAPH LEARNING WITH SPARSITY REGULARIZATION

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

[1] Mitsubishi Electric Research Labs (MERL),
201 Broadway, Cambridge, MA 02139, USA
[2] Department of Electrical Engineering, University of Southern California,
3740 McClintock Ave., Los Angeles, CA 90089, USA

ABSTRACT

Learning graph topology from data is challenging. Previous work leads to learning graphs on which the graph signals used for training are smooth. In this paper, we propose an optimization framework for learning multiple graphs, each associated to a class of signals, such that representation of signals within a class and discrimination of signals in different classes are both taken into consideration. A Fisher-LDA-like term is included in the optimization objective function in addition to the conventional Gaussian ML objective. A block coordinate descent algorithm is then developed to estimate optimal graphs for different categories of signals, which are then used to efficiently classify the different signals. Experiments on synthetic data demonstrate that our proposed method can achieve better discrimination between the learned graphs, leading to improvements in subsequent classification tasks.

***Index Terms*—** Graph learning, precision matrix estimation, graph signal processing, discriminant analysis, classification

1. INTRODUCTION

Graph structures arise naturally in various domains that require manipulating structured data, such as sensor networks, image and video coding/processing, geographic data and social networks [1][2]. In graph signal processing, the signal is defined as a function on the vertex set of the graph so that classical signal processing techniques can be extended to this irregular domain. Typically, the vertices represent data entities and the edges represent the pairwise relationships between them. However, constructing the optimal graph, i.e., selecting edges and the corresponding weights, is not trivial. In some cases, there is a clear choice for graphs based on prior or domain-specific knowledge, e.g., 4-connected graph is popular in image coding. However, in other applications, intuitive choices for graphs may not always reflect the real intrinsic relationships between entities. Hence, learning the right graph topology from observed data is an important research topic in graph signal processing.

Previous research tackles the graph learning problem from various perspectives, which are often based on promoting smoothness of data samples on the learned graph. For example, in [3], the graph is learned by solving an optimization problem, where the objective function includes two terms: one is for measuring the smoothness on the noiseless version of observed graph signals, and the other is a data fitting term. Such objective functions are selected in order to favor efficient signal representation, so that when representing signals in terms of the graph spectrum (i.e., in the graph Fourier transform domain) a small number of non zero coefficients is sufficient (on average) to provide a good approximation. This is desirable for applications such as denoising and compression.

Another family of graph learning approaches considers the problem from a probabilistic perspective, with graph signals viewed as random vectors with a Gaussian Markov Random Field (GMRF) distribution and the precision matrix playing the role of the Laplacian matrix [4]. Under this framework, learning the graph topology and associated graph Laplacian, i.e., estimating the precision matrix, can be formulated as solving a Gaussian Maximum Likelihood (ML) problem with an additional ℓ_1 regularization term, which encourages learning a sparse graph [5][6][7]. In [5] and [6], a coordinate descent procedure is applied to efficiently solve the ℓ_1-penalized Gaussian ML problem. Notice that solving the Gaussian ML problem will yield a trace term that can also be interpreted as a smoothness measure, leading to energy compaction of signals on the learned graphs. In addition, including the ℓ_1-penalty leads to a learned graph with sparser connectivity, which can make it easier to interpret statistical dependencies present in the data.

In this paper, we focus on applications involving classification. A limited number of data samples are available, where each sample is associated with one known category/label. A class-specific sub-dictionary can be defined, e.g., by learning from those data samples in the corresponding class. Cascading these sub-dictionaries of each class leads to one dictionary, which can then be used to sparsely code each data sample. The resulting sparse coding coefficients can serve as the features to any conventional classifier, e.g., kernelized SVM, in order to obtain the labels for input data.

Under the above-mentioned scenario, how to construct the class-specific sub-dictionaries is critical to the classification performance. In a recent series of works [8][9][10], one common graph is defined for all the categories and class-specific graph transforms, i.e., polynomials of defined graph Laplacian, are learned based on signals in each category, which then act as the sub-dictionaries. Using one common graph for all the categories will lead to a lack of discrimination between the class-specific graph transforms, which degrades the classification performance. One may consider applying the graph learning approaches mentioned above, such as graphical lasso, independently to each class of signals (each category), in order to learn one graph for each class. However, since those methods design graphs that will favor energy compaction and sparsity within each class, utilizing the resulting graph transform on each learned graph as sub-dictionary does not guarantee that it will be effective in

This work was done when Jiun-Yu Kao was an intern at MERL.

Work supported in part by a grant from Mitsubishi Electric. Antonio Ortega also worked as a consultant for Mitsubishi Electric Research Labs (MERL).

discriminating between classes.

To address this problem, we propose to take discrimination into account when constructing class-specific graphs from signals, leading to discriminative class-specific sub-dictionaries. Thus, each learned graph will take into consideration not only energy compaction of signals in that class, but also discrimination with respect to signals in other classes. In this way, the resulting graph transforms, e.g., GFT basis, will be able to discriminate between classes and will be better suitable for classification. To the best of our knowledge, we are the first to propose a graph estimation method that combines both a fitting term to optimize energy compaction and a term to promote discrimination.

The rest of the paper is organized as follows. In Section 2, we formally define the multiple-category graph learning problem, along with the proposed objective function to be optimized. In Section 3, we derive the block coordinate descent algorithm for solving the proposed optimization problem. The experimental results on synthetic data and related discussions are presented in Section 4. We discuss conclusions and potential extensions for future work in Section 5.

2. PROBLEM FORMULATION

Inspired by the Fisher discrimination criterion for linear discriminant analysis (LDA) [11], which aims at minimizing the within-class scatter while maximizing the between-class scatter, we propose a Fisher discrimination graph learning algorithm where the graph representing each signal category is learned jointly from data samples in all the categories. Based on the conventional Gaussian ML objective to estimate the graph Laplacian for each category [6], a new objective function is proposed that includes an additional term measuring the non-smoothness of data from other categories on the graph of a specific category. This object function favors smoothness within a class as well as non-smoothness across classes, with the goal to improve the discrimination among the learned graphs for different classes.

We first define the multi-category graph learning problem where we are interested in learning multiple graphs, each for one category of signals (data samples). Assume there are n-dimensional random graph signals, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$, where $\mathbf{x}^{(i)}$ is the random signal associated with i-th category (label) and there are S categories in total. Furthermore, for each category i of signals, there are N_i- i.i.d. realizations, $\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_{N_i}^{(i)}$. Note that $\mathbf{x}_j^{(i)} \in \mathbb{R}^{n\times 1}$.

Our goal is to learn the graph structure of each category, i.e. $\mathcal{G}_1, \ldots, \mathcal{G}_S$, from the observed random signals. Each weighted undirected graph $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}_i, \mathbf{Q}_i)$ consists of a set of vertices $\mathcal{V} = \{1, 2, \ldots, n\}$ connected by a set of edges $\mathcal{E}_i$ and a symmetric matrix representation $\mathbf{Q}_i$, where for $a \neq b$, $(a, b) \in \mathcal{E}_i$ if and only if $\mathbf{Q}_{i,ab} \neq 0$. From the graph signal processing literature, $\mathbf{Q}_i$ has usually been restricted to a graph Laplacian [1], a generalized graph Laplacian [12] or an adjacency matrix [2]. Here we consider the least restricted constraint for $\mathbf{Q}_i$ where it is only required to be positive semi-definite, similar to the graphical lasso method [5].

Applying the conventional graphical lasso method directly to this multi-category graph learning setting leads to solving the following ℓ_1-penalized Gaussian ML estimation problem separately for the graph of each category i,

$$\min_{\mathbf{Q}_i \succeq 0} \; -\log\det(\mathbf{Q}_i) + \mathrm{tr}(\mathbf{K}_i\mathbf{Q}_i) + \rho\|\mathbf{Q}_i\|_1, \tag{1}$$

where $\mathbf{K}_i$ represents the empirical covariance matrix computed from the realizations $\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_{N_i}^{(i)}$.

By revisiting the trace term in (1),

$$\mathrm{tr}(\mathbf{K}_i\mathbf{Q}_i) = \frac{1}{N_i}\mathrm{tr}\left(\sum_{k=1}^{N_i}\mathbf{x}_k^{(i)}\mathbf{x}_k^{(i)^T}\mathbf{Q}_i\right) = \frac{1}{N_i}\sum_{k=1}^{N_i}\mathbf{x}_k^{(i)^T}\mathbf{Q}_i\mathbf{x}_k^{(i)}, \tag{2}$$

it can be observed that minimizing $\mathrm{tr}(\mathbf{K}_i\mathbf{Q}_i)$ is promoting the average smoothness of the realizations (data samples) in the i-th category on the estimated i-th graph.

In order to promote discrimination between graphs learned with signals in different categories, it would be desirable for the signals in i-th category not only to be smooth on the graph learned for the i-th category, but also be non-smooth on the learned graphs corresponding to other categories. To address this desired property, we propose to modify the trace term in (1) as follows:

$$\begin{aligned}&\min_{\mathbf{Q}_1,\ldots,\mathbf{Q}_S} \sum_{i=1}^{S}\sum_{k=1}^{N_i}\frac{1}{N_i}\left[\mathbf{x}_k^{(i)^T}\mathbf{Q}_i\mathbf{x}_k^{(i)} - \frac{1}{S-1}\sum_{j\neq i}^{S}\mathbf{x}_k^{(i)^T}\mathbf{Q}_j\mathbf{x}_k^{(i)}\right]\\ =&\min_{\mathbf{Q}_1,\ldots,\mathbf{Q}_S} \sum_{i=1}^{S}\left[\mathrm{tr}\left(\mathbf{K}_i\mathbf{Q}_i\right) - \frac{1}{S-1}\sum_{j\neq i}^{S}\mathrm{tr}\left(\mathbf{K}_i\mathbf{Q}_j\right)\right],\end{aligned} \tag{3}$$

where $\mathbf{K}_1, \ldots, \mathbf{K}_S$ are the empirical covariance matrices of signals from each category.

After reformulating (3), we propose to solve the following new optimization problem,

$$\min_{\mathbf{Q}_i\succeq 0} -\log\det(\mathbf{Q}_i) + \mathrm{tr}(\mathbf{K}_i\mathbf{Q}_i) - \frac{\mu_i}{S-1}\sum_{j\neq i}^{S}\mathrm{tr}(\mathbf{K}_j\mathbf{Q}_i) + \rho_i\|\mathbf{Q}_i\|_1, \tag{4}$$

for each $\mathbf{Q}_i$, given $\mathbf{K}_1, \ldots, \mathbf{K}_S$. μ_i and ρ_i represent the weight for each regularizer. For simplicity, we assume $\mu_i = \mu$ and $\rho_i = \rho$ for all i, in the remainder of this paper.

3. DISC-GLASSO ALGORITHM

In this section, we develop a block coordinate descent algorithm similar to that in [5] for solving the optimization problem (4).

First the subgradient of (4) is

$$-\mathbf{Q}_i^{-1} + \mathbf{K}_i - \frac{\mu}{S-1}\sum_{j\neq i}^{S}\mathbf{K}_j + \rho\mathbf{\Gamma}_i = 0, \tag{5}$$

where $\mathbf{\Gamma}_i = \mathrm{sign}(\mathbf{Q}_i)$. Letting $\frac{\mu}{S-1} = \frac{1}{r}$ we can rewrite (5) as:

$$-\mathbf{W}_i + \mathbf{K}_i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j + \rho\mathbf{\Gamma}_i = 0 \tag{6}$$

where $\mathbf{W}_i$ is the estimated covariance matrix and $\mathbf{W}_i = \mathbf{Q}_i^{-1}$.

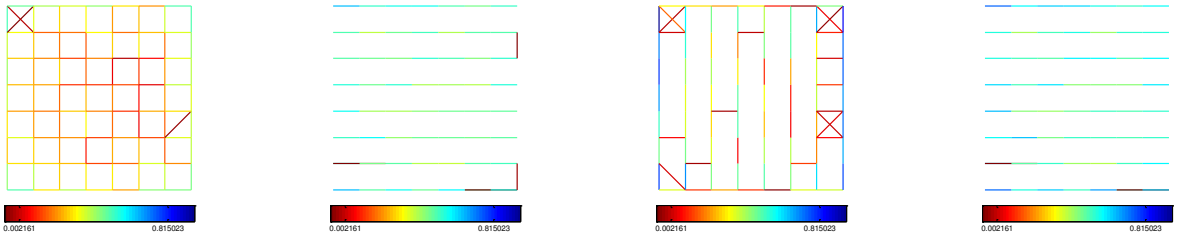
Consider a partition for $\mathbf{W}_i$ and $\mathbf{K}_i$,

$$\mathbf{W}_i = \begin{pmatrix}\mathbf{W}_{11}^i & \mathbf{w}_{12}^i\\ \mathbf{w}_{12}^{i\,T} & w_{22}^i\end{pmatrix}, \mathbf{K} = \begin{pmatrix}\mathbf{K}_{11}^i & \mathbf{k}_{12}^i\\ \mathbf{k}_{12}^{i\,T} & k_{22}^i\end{pmatrix} \tag{7}$$

where $\mathbf{W}_{11}^i, \mathbf{K}_{11}^i$ are $(n-1)\times(n-1)$ sub-matrices, $\mathbf{w}_{12}^i, \mathbf{k}_{12}^i$ are column vectors of length $n-1$, and use similar partitions for $\mathbf{\Gamma}_i$.

The upper right block of (6) leads to

$$-\mathbf{w}_{12}^i + \mathbf{k}_{12}^i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{k}_{12}^j + \rho\gamma_{12}^i = 0 \tag{8}$$



(a) GLasso, Category 1, $|\mathcal{E}| = 115$

(b) GLasso, Category 2, $|\mathcal{E}| = 60$

(c) Disc-GLasso, Category 1, $|\mathcal{E}| = 88$

(d) Disc-GLasso, Category 2, $|\mathcal{E}| = 58$

Fig. 1: Visualize the learned graphs for two categories of graph signals with different graph learning methods.

Furthermore, solving (8) will be equivalent to solving the following dual problem:

$$\beta = \arg\min \frac{1}{2} \left\| \mathbf{W}_{11}^{i\ 1/2}\beta - \mathbf{W}_{11}^{i\ -1/2}\mathbf{k}_{12}^{i} + \frac{1}{r}\sum_{j\neq i}^{S} \mathbf{W}_{11}^{i\ -1/2}\mathbf{k}_{12}^{j} \right\|^2 + \rho\|\beta\|_1, \tag{9}$$

as once β solves (9), $\mathbf{w}_{12}^{i} = \mathbf{W}_{11}^{i}\beta$ can solve (8).

The above procedure can then be repeated through all the columns/rows partition until convergence has been reached, as shown in Algorithm 1. Compared to [5], our algorithm takes into consideration the partitioned weighted covariance matrices of other classes, in addition to that of the target class. This makes initialization of our algorithm more challenging since we need to set the parameter r appropriately; we discuss this next.

Criterion for choosing r:

To guarantee that after each updating step t, $\mathbf{W}_i^{(t)} \succ 0, \forall t$, we show below that only the initial $\mathbf{W}_i$ is required to be positive definite: $\mathbf{W}^{(0)} \succ 0$. As the initialization is defined as $\mathbf{W}_i^{(0)} := \left(\mathbf{K}_i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j\right) + \rho\mathbf{I}$, we need to have

$$\begin{aligned} &\mathbf{K}_i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j \succeq 0 \\ &\rho > 0 \end{aligned} \tag{10}$$

Choosing r so that (10) is satisfied can lead to the corresponding $\mathbf{W}^{(0)} \succ 0$. Now suppose that $\mathbf{W}^{(t)} \succ 0$, which implies the Schur complement is positive: $w_{22} - \mathbf{w}_{12}^T(\mathbf{W}_{11}^{(t)})^{-1}\mathbf{w}_{12} > 0$. Then by the update rule, the corresponding Schur complement for updated $\mathbf{W}^{(t+1)}$ will be even greater:

$$w_{22} - \mathbf{w}_{12}^T(\mathbf{W}_{11}^{(t+1)})^{-1}\mathbf{w}_{12} > w_{22} - \mathbf{w}_{12}^T(\mathbf{W}_{11}^{(t)})^{-1}\mathbf{w}_{12} > 0$$

Thus, once $\mathbf{W}^{(0)} \succ 0$, the consecutive updated $\mathbf{W}^{(t)} \succ 0, \forall t$.

Therefore, in the proposed algorithm, we first search for the best (minimum) ratio r such that $\mathbf{K}_i - \frac{1}{r}\sum_{j\neq i}^{S}\mathbf{K}_j \succeq 0, \forall i$, via line search through a predefined set of possible values for r. Compared to GLasso [5], the only additional cost in time complexity of adding the proposed Fisher-LDA-like term is this searching procedure.

4. EXPERIMENTS

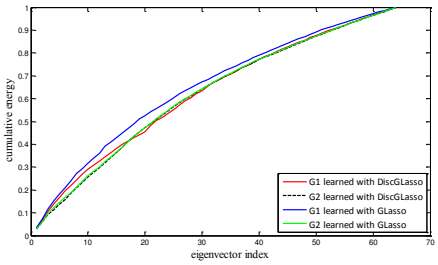
In this section, we generate synthetic data to validate that our proposed method can construct graphs that are better at discriminating between multiple categories. First, we construct two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, each having 64 vertices following the 8×8 grid pattern. $\mathcal{G}_1$ is a 4-connected graph with equally weighted horizontal and vertical edges ($w_{hor} = w_{ver} = 0.9$) while $\mathcal{G}_2$ is also 4-connected but with heavily weighted horizontal edges and weakly connected vertical edges ($w_{hor} = 0.9, w_{ver} = 0.1$). The combinatorial Laplacian matrices, i.e., $\mathbf{L}_1$, $\mathbf{L}_2$, are then constructed for each graph and $\mathbf{K}_1 = (\mathbf{L}_1 + \sigma_\epsilon\mathbf{I})^{-1}$, $\mathbf{K}_2 = (\mathbf{L}_2 + \sigma_\epsilon\mathbf{I})^{-1}$ are computed. Then we generate $N = 2000$ i.i.d. realizations following multivariate Gaussian distribution with covariance matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ respectively for each class. We refer to these 4000 graph signals as training data samples for learning the graph topology of each category. Finally the empirical covariance matrices computed from training samples in two categories are used as input to our proposed Disc-GLasso algorithm, with $\sigma_\epsilon = 1.0$ and $\rho = 0.05$. As for comparison, conventional graphical lasso [5] is applied on the empirical covariance matrix of training samples in each class to estimate the graph of that category respectively.

The learned graphs for each data category and with each method are visualized in Fig. 1. Qualitative results show that, instead of pursuing solely smoothness/energy compaction, the graph for signals in category 1 that have been learned with the proposed Disc-GLasso algorithm have more strongly connected vertical edges, which provide better discrimination between these two graphical models, while the one learned with conventional graphical lasso has more equally-weighted horizontal and vertical edges.

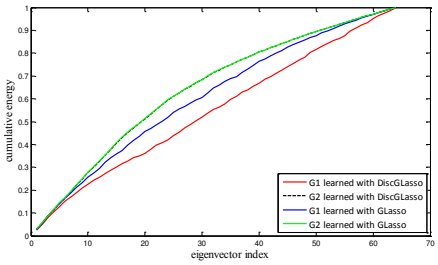
Algorithm 1: Disc-GLasso algorithm

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

1. Search for the best (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. Cycle through the columns repeatedly, performing following steps till convergence:
   (a) Rearrange the rows/columns so that the target column is last (implicitly).
   (b) Solve the lasso problem stated in (9).
   (c) Fill in the corresponding row and column of $\mathbf{W}_i$ using $\mathbf{w}_{12}^i = \mathbf{W}_{11}^i\hat{\beta}$.



(a) test signals in category 1



(b) test signals in category 2

Fig. 2: Cumulative spectrum energy of test signals in each class on the learned graphs. G1 and G2 represent respectively the graph learned for each category.

As for the quantitative experiments, we generate another N i.i.d. Gaussian samples for each class with the same covariance matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ to represent the testing data samples. The following three measures are adopted to validate the improved discriminability.

- The cumulative spectrum of signals in each category on the graphs learned via Disc-GLasso and GLasso.
- The self-defined separation measure:

$$s = \frac{\mathrm{tr}(\mathbf{K}_1\mathbf{Q}_2) + \mathrm{tr}(\mathbf{K}_2\mathbf{Q}_1)}{\mathrm{tr}(\mathbf{K}_1\mathbf{Q}_1) + \mathrm{tr}(\mathbf{K}_2\mathbf{Q}_2)} \tag{11}$$

- The classification accuracy between two categories of signals.

The cumulative spectrum of signals on the learned graphs are plotted in Fig. 2. As expected, the test signals in category 2 are shown to be smooth on the graph learnt for class 2, regardless of whether they are learned with GLasso or Disc-GLasso. Furthermore, they are much less smooth on the graph of class 1 learned with proposed Disc-GLasso than on that with GLasso, which validates the discriminative power between graphs.

Fig. 3 plots the separation measure as defined in (11) versus the parameter r. When discriminant graphs learned for each class, the denominator of s will be small while the numerator will be large, as the signals of one class should be smooth on the learned graph of that class and not smooth on the graph of the other class, which leads to a large s. Notice that as r increases, the effect of the additional term $\sum_{j\neq i}^{S} \mathrm{tr}(\mathbf{K}_j\mathbf{Q}_i)$ becomes weaker, leading to less discrimination between the learned graphs, which is again validated via the monotonically descending trend in Fig 3.

Finally, we examine whether improving the discrimination between learned graphs of different classes translates to improved classification. The classification accuracy between the two categories is reported in Fig. 4. Only the training samples are utilized for training the classifier and the accuracy is reported based on the 4000 testing signals. It is worth noting that our method can be utilized to

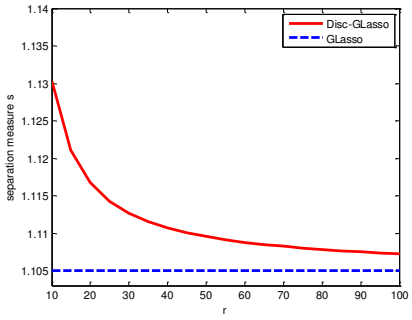


Fig. 3: The separation measure versus r with Disc-GLasso compared to that of GLasso.

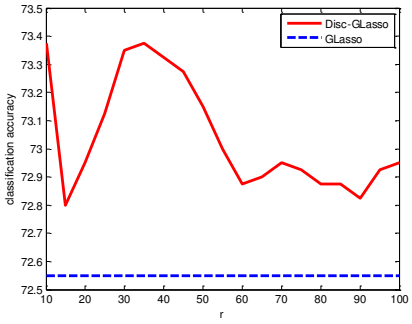


Fig. 4: Classification accuracy versus r with Disc-GLasso compared to that of GLasso.

preprocess the signals before any conventional classifiers, such as SVM, is applied. The classification scheme we apply in this experiment directly assigns a class label to each test signal by choosing the class such that the corresponding graph transform provides a more compact representation of the input signal. Specifically, we project each test signal onto the first $\frac{1}{2}$ low-frequency GFT basis computed based on graph learned for each category and calculate the projection energy. Then the label of each signal is assigned with the class whose low-frequency GFT basis preserve more energy. The classification accuracy shows a consistent improvement versus selected ratio r when the graphs are learned with our proposed method.

5. CONCLUSIONS

We propose a novel graph learning approach that learns a discriminative set of graphs from multiple categories of data samples. Instead of learning the graphs in terms of representability, we propose to include an additional LDA-like term to enable a better discriminability between classes. We also derive a block coordinate descent algorithm to efficiently estimate the graph topology of each class. Qualitative and quantitative experiments on a synthetic dataset demonstrate that the graphs learned with our proposed method have more discrimination between classes, leading to benefits for classification. Future work will consider applying our method on real-world data, such as anomaly detection, and also considering other functionals that may further improve discriminability between graphs.

6. REFERENCES

[1] D. I Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[2] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 80–90, 2014.

[3] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Laplacian matrix learning for smooth graph signal representation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 3736–3740.

[4] P. A. Chou C. Zhang, D. Florencio, "Graph signal processing - a probabilistic framework," Tech. Rep., 2015.

[5] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[6] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electronic journal of statistics*, vol. 6, pp. 2125, 2012.

[7] B. M. Lake and J. B. Tenenbaum, "Discovering structure by learning sparse graphs," *Proceedings of the 32nd Cognitive Science Conference*, 2010.

[8] Xuan Zhang, Xiaowen Dong, and Pascal Frossard, "Learning of structured graph dictionaries," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[9] D. Thanou, D. I. Shuman, and P. Frossard, "Parametric dictionary learning for graph signals," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, 2013, pp. 487–490.

[10] D. Thanou, D. I. Shuman, and P. Frossard, "Learning parametric dictionaries for signals on graphs," *IEEE Transactions on Signal Processing*, vol. 62, no. 15, pp. 3849–3862, 2014.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning : data mining, inference, and prediction (Second Edition)*, Springer, 2009.

[12] E. Pavez and A. Ortega, "Generalized laplacian precision matrix estimation for graph signal processing," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6350–6354.