Sparse Representation-Based Classification of Attributed Graphs

raph powpow
raphael.phong@gmail.com

***Abstract*—Experiments still running, so no abstract. :-)**

***Index Terms*—Dictionary learning, graph classification, graph representation learning, graph signal processing, sparse coding**

I. Introduction

The graph continues to enjoy both economical and mathematical interest as the simplest topological structure over which signals can be defined, manipulated, and – hopefully – interpreted. Of great interest and yet of notorious difficulty is the fairly new task of classifying graphs endowed with an informative, node-wise signal (from here on referred to as "attributed graphs"). Tools previously developed by the signal processing community like spectral filtering [2], [3], wavelet transforms [1], sampling algorithms [4] and so on allow for a rigorous treatment of nodal signals, albeit in the limited context where only one graph is considered. Developped adjacently, graph neural networks – conceived from the paradigms of parameterized message propagation [5], [8] and learned filtering [6] – offer a heavy-duty solution for attributed graph classification within the setting of multiple graphs. Still, one notable challenge of this task lies in the "reading out" of each graph. It is not evident to devise a uniform procedure that takes an attributed graph and returns a summary of its global information, because the number of nodes (i.e., the signal length) and the domain topology varies from graph to graph. One popular remedy is the graph pooling framework [7], which essentially performs recursive subsampling on the graph (often interlaced with the message propagation/filter layers) followed by the application of a global, permutation-invariant function ϕ to each graph subsample (e.g. a signal mean across the remaining nodes); the results are combined with some operation $\oplus$ (e.g. summation or concatenation) to give the graph readout [8]. However, a common issue lies in the summative nature of ϕ, whereby loss of local information is unavoidable. Moreover, the pooling operation often requires an intricate reconstructive step where new edges need to be assigned between the remaining nodes after subsampling, hence requiring a supplementary learnable pair-wise operator to be iterated over all pairs of nodes.

In light of these inconveniences, we revisit the approaches of sparse representation-based classification [10], employing a bank of spectral kernels to sparsely encode signals across graphs with the hopes of providing an alternative to pooling for graph classification. After having bridged the gap between the traditional time-domain and the nodal domain with regards to dictionary learning theory in Sections III and IV, we provide in Section V a model which uses the same discriminative penalties as those of [11]. Data processing, training and experiment details are presented in Sections VI and VII. Section II provides preliminary theory.

II. Related work

A. Dictionary Learning for Classification

First, an early approach to the problem is to define one dictionary for each signal class, in the hopes that the dictionary yielding the sparsest, best reconstruction of a signal f corresponds to the true label of f. Such works include [10], [12], [13]; [12] adds a softmax penalty to endow the dictionaries with discriminative power.

Another approach is to apply a linear classifier to the sparse coefficients [14], [18]. Joint optimization of both the classifier and dictionaries are proposed in [15]. Ref. [16] moreover adds a label consistent penalty term which enforces an association between dictionary atoms and particular labels. Other methods of discrimination within this framework include the use of Fisher discrimination within the representation coefficients [17], [18]. Ref. [19] separately learns one sub-dictionary for each class, removes the redundant atoms shared between them, and then combines them all to form one dictionary. Inspired by the latter, [11] continues to enforce an atom-to-class association while distinguishing discriminating atoms from common atoms.

B. Dictionary Learning on Graphs

There already exists literature which extends the framework of dictionary learning to the case of graph signals, albeit without a focus on signal classification. To wit, Zhang et al. [20] devise a dictionary composed of multiple sub-dictionaries, each obtained by the modulation of the graph Laplacian's eigenvalues – this approach does not lend itself to the setting where signals may lie on different graphs. On the other hand, Thanou et al. [21] use learnable polynomials of the Laplacian to construct the sub-dictionaries. This approach is amenable to the "multiple graph" setting because dictionaries are constructed from kernels defined on the frequency domain common to all graphs. However, the authors impose a set of constraints (e.g. frame bounds [22] and smoothness) which ensure, in general, a good reconstruction quality. Since we

really seek to discriminate across signal labels using the reconstruction errors, consistently good dictionaries are undesirable and so we forego these additional constraints.

It is important to note that the above dictionaries are all "convolutional" in nature [23]: since they are assembled from translated kernels, the atoms enjoy more localization, although incurring higher computational cost and possibly adding redundancy in the representation atoms. We instead propose to use the untranslated kernel, at the cost of having to diagonalize each graph Laplacian as a pre-processing step.

III. Preliminaries

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with nodes $\mathcal{V}$ (say, N of them) and edges $\mathcal{E}$; also, we keep an implicit ordering on the nodes. If $\mathcal{A}$ and D are respectively the adjacency and the degree matrix of $\mathcal{G}$, then the normalized Laplacian of $\mathcal{G}$ is given by $\mathcal{L} = I - D^{-1/2}\mathcal{A}D^{-1/2}$. The Laplacian is positive-semidefinite, and as such always admits an eigendecomposition $\mathcal{G} = U\Lambda U^T$, with Fourier modes $U = [\chi_0, ..., \chi_{N-1}]$ and frequencies $\Lambda = \text{diag}(\lambda_0, ..., \lambda_{N-1})$. It follows that the graph Fourier transform of some $f : \mathcal{V} \to \mathbb{R}$ is given by $\hat{f} = U^T f$, while the inverse of $\hat{f}$ is given by $U\hat{f} = f$. Given a kernel $g : [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}$ over the spectrum, its nodal Fourier inverse is given by $\psi_g = U[g(\lambda_0), ..., g(\lambda_{N-1})]^T$, and its translation by node $i \in \mathcal{V}$ is given by a (normalized) convolution with the Dirac delta δ_i:

$$\psi_{g,i} := \sqrt{N} U g(\Lambda) U^T \delta_i.$$

As mentionned in [24], this definition of a convolution works backward from the notion that the graph Fourier transform should diagonalize the "node-domain" convolution $(\psi_g * \delta_i)$, but it is indeed well defined ([24], Section 5.3). More generally, $\sqrt{N} U g(\Lambda) U^T$ is the matrix $\sqrt{N} g(\mathcal{L})$, and the columns correspond to individual translates of the graph kernel. If g is a polynomial, then an intrusive diagonalization is not needed ([1], Section 6)

Thanou et al. build their dictionary by assembling the combined matrix $D = [Ug_1(\Lambda)U^T, ..., Ug_M(\Lambda)U^T]$. In general, when g is a smooth kernel with uniformly bounded derivatives, the magnitude of $\psi_{g,i}(n)$ will enjoy a quantifiable decay as the hop-distance between nodes i and n increases ([24], Corollary 2). This localization is desirable when one wants to perform a windowed convolution on some signal f with the kernel g, and may contribute positively to a better signal reconstruction. However, as previously mentioned, we want finer control over the dictionary where, ideally, each atom has its own set of parameters. The goal is to allow for an easier enforcement of discriminative capabilities, which is prioritized over reconstructive capabilities. To this end, we construct our dictionaries with the untranslated kernels ψ_g.

IV. Dictionary learning on graph signals

A. Block-matrix formulation

In practice, graphs may support signals with C different channels, as is the case when each node is assigned a vectorial embedding [25]. The graph signal is then actually a matrix $F = [f_1, ..., f_C] \in \mathbb{R}^{N \times C}$. The goal, then, is to construct a learnable dictionary composed of C-channel atoms, say M of them. To this end, we outline an approach highly similar to Thanou et al., but the construction of our dictionary is different in that we assemble a block matrix of the form

$$D = \begin{bmatrix} \psi_{g_{1,1}} & \cdots & \psi_{g_{1,M}} \\ \vdots & \ddots & \vdots \\ \psi_{g_{C,1}} & \cdots & \psi_{g_{C,M}} \end{bmatrix}. \tag{1}$$

The graph signal to be reconstructed then corresponds to the flattened node embeddings $f = [f_1^T, ..., f_C^T]^T$. Finally, one optimizes over $\mathbb{R}^M$ in search of the best atomic representation r^* such that

$$\min_{r \in \mathbb{R}^M} \frac{1}{2}||f - Dr||_2^2 + \lambda||r||_1, \tag{2}$$

is attained. Here, $\lambda > 0$ controls the enforcement of sparsity. One can also optimize over the kernel bank when the kernels have explicit parameterization, e.g. when they are polynomials. Consider the joint problem

$$\min_{\substack{r \in \mathbb{R}^M \\ \alpha_{1,1}, ..., \alpha_{C,M} \in \mathbb{R}^K}} \frac{1}{2}||f - Dr||_2^2 + \lambda||r||_1, \tag{3}$$

where $\alpha_{i,j}$ is a vector containing K polynomial coefficients such that $g_{i,j}(x) = \sum_{0 \le k \le K} \alpha_{i,j}(k) x^k$.

B. QCQP reformulation

Problem (3) is convex in $A = (\alpha_{1,1}, ..., \alpha_{C,M})$ when r is fixed, and vice-versa. For clarity of exposition, we derive a reformulation of Problem (3) which details the nature of A as an optimization variable. First, split the objective along the channels to get

$$\min_{A \in \mathbb{R}^{K \times CM}} \frac{1}{2} \sum_{1 \le i \le C} ||f_i - [\psi_{g_{i,1}}, ..., \psi_{g_{i,M}}]r||_2^2. \tag{4}$$

Now, define the vector γ^k by $\gamma^k = [\lambda_0^k, ..., \lambda_{N-1}^k]^T$ and let $\Gamma_K = [\gamma^0, ..., \gamma^K]$. Observe that

$$\psi_{g_{i,j}} = U g_{i,j}(\gamma^1) = U \left(\sum_{0 \le k \le K} \alpha_{i,j}^{(k)} \gamma^k \right) = U \Gamma_K \alpha_{i,j}.$$

In other words, letting $\alpha_i = [\alpha_{i,1}^T, ..., \alpha_{i,M}^T]^T$, the objective can be rewritten as a sum of independent quadratic programs:

$$\min_{\alpha_1, ..., \alpha_C \in \mathbb{R}^{KM}} \frac{1}{2} \sum_{1 \le i \le C} ||f_i - [r(1)U\Gamma_K, ..., r(M)U\Gamma_K]\alpha_i||_2^2 \tag{5}$$

The convexity of Problem (5) is useful when fitting A over the entire training set at once is feasible. A closed form update rule can even be derived by zeroing the Sample Average Approximated (SAA) gradient [26]. Still, the overall optimizing task is non-convex w.r.t. (A, r) and so we would still need to employ an alternating optimization scheme going back and forth between A and r [27]. For future reference, denote $R_x := [x(1)U\Gamma_K, ..., x(M)U\Gamma_K]$ and $D_k = [\psi_{g_{k,1}}, ..., \psi_{g_{k,M}}]$.

V. Graph classification model

Yang et al. [17] and Kong et al. [11] both construct a classification model by partitioning a common dictionary into L specialized sub-dictionaries, where L is the number of classes. Their approaches are especially similar with regards to how the objective function is conceived. We choose to adapt the latter model for its unique consideration of common, indiscriminate reconstructive atoms.

Let f belong to class i, $1 \leq i \leq L$ and let the number of atoms $M = m(L+1)$ be a multiple of $L+1$. Partition the dictionary D and the representation r into $L+1$ components so that $D = [D_1, ..., D_{L+1}]$ and $r = [r_1^T, ..., r_{L+1}^T]^T$. Here, only the sub-dictionaries D_i and D_{L+1} should be crucial for the reconstruction of f. The atoms of D_{L+1} are expected to provide an essential reconstructive basis common to all classes, hence lifting any burden of representational redundancy for the class-specific sub-dictionaries $D_1, ..., D_L$ and encouraging specialization. We wish to minimize

$$O(D,r) = \frac{1}{2}\Bigg(||f - Dr||_2^2 + \sum_{\substack{1 \leq j \leq L \\ j \neq i}} ||r_j||_2^2 + ||f - [D_i, D_{L+1}] \begin{pmatrix} r_i \\ r_{L+1} \end{pmatrix} ||_2^2 \Bigg) + \lambda ||r||_1 + \eta \sum_{\substack{1 \leq j \leq L+1 \\ j \neq i}} ||\bar{D}_i^T \bar{D}_j||_F^2$$

Here, $\bar{D}_i$ denotes the column-normalized version of D_i. Note that, as in [11], the original reconstruction and sparsity terms in (2) are preserved. The incoherence term $\sum_{1 \leq j \leq L+1, j \neq i} ||\bar{D}_i^T \bar{D}_j||_F^2$ scaled by $\eta > 0$ further enforces dissimilarity between the sub-dictionaries. As formulated by Kong et al., one can rewrite the above objective as

$$O(D,r) = \frac{1}{2} \left|\left| \begin{pmatrix} f \\ f \\ 0 \end{pmatrix} - \begin{pmatrix} D \\ D\tilde{Q}_i \tilde{Q}_i^T \\ Q_{/i}^T \end{pmatrix} r \right|\right|_2^2 + \lambda ||r||_1 + \eta \sum_{\substack{1 \leq j \leq L+1, j \neq i}} ||\bar{D}_i^T \bar{D}_j||_F^2 \tag{6}$$

using the selection operators defined by $Q_i = [q_1^i, ..., q_p^i]$ and $\tilde{Q}_i = [Q_i, Q_{L+1}]$, with q_j^i, $1 \leq j \leq p$, the vector in $\mathbb{R}^M$ that is zero everywhere except the $(ip+j)$th entry and $Q_{/i} = [Q_1, ..., Q_{i-1}, Q_{i+1}, ..., Q_L]$.

A. Full batch dictionary updates

The incoherence term of (6) is non-convex with respect to A. We provide here an alternative convex penalty which still encourages a form of between-class decorrelation. As a result, a fully convex objective is obtained, which allows for efficient full batch dictionary updates. One small requirement of this adjustment is that we now work exclusively with adjacency matrices rather than Laplacians.

The normalized eigenvalues of all sample adjacency matrices from class ℓ (group them into the set Λ_ℓ) are used to estimate a hypothesized spectral density[1] p_ℓ [] with the discrete distribution $(1/|\Lambda_\ell|) \sum_{\lambda \in \Lambda_\ell} \delta(\lambda - \cdot)$. For computational feasibility, a further estimation is made by quantizing each class's discrete distribution into a uniformly-binned histogram over $[\lambda_{\min}, \lambda_{\max}]$; the result is a set H consisting of all bin midpoints, giving a weighted sum of much fewer Dirac deltas $(1/|\Lambda_\ell|) \sum_{\lambda \in H} N_{\ell,\lambda} \delta(\lambda - \cdot)$, $\sum N_{\ell,\lambda} = |\Lambda_\ell|$ as an estimate of p_ℓ. A spectral kernel $g_{i,j}$ (with coefficients $\alpha_{i,j}$) associated to label ℓ is expected to be less "active" where the eigenvalues of some label $\ell-$graph are less likely to fall than those from another label's sample. To this end, we seek to minimize the absolute overlap of order p

$$\mathcal{Q}_\ell^p(g_{i,j}) := \frac{1}{L-1} \sum_{\lambda \in H} |\omega_{\ell,\lambda} g_{i,j}(\lambda)|^p \approx \int_{[\lambda_{\min}, \lambda_{\max}]} |q_{/\ell}(x) g_{i,j}(x)|^p dx \tag{7}$$

between the kernel and the heuristic weighting factor $q_{/\ell}(x) := \frac{p_{/\ell}(x)}{p_\ell(x)}$, where $p_{/\ell}(x) = \frac{1}{L-1} \sum_{k \neq \ell} p_k(x)$ is the density corresponding to the process of sampling an eigenvalue from the density p_k of any other class chosen randomly. The (unnormalised) bin weights are given by $\omega_{\ell,\lambda} = \frac{|\Lambda_\ell|}{N_{\ell,\lambda}} \sum_{k \neq \ell} \frac{N_{k,\lambda}}{|\Lambda_k|}$. For future reference, denote $\Omega_{/\ell} = \text{diag}(\omega_{\lambda_1}, ..., \omega_{\lambda_{|H|}})$.

Consider now the full-batch update of A in the case of channel $k \in \{1, ..., C\}$. Returning to the notation of Section IV, rewrite the first term of (6) (dropping the discriminative L^2 penalty on r) as a standard quadratic program:

$$\begin{aligned} &\frac{1}{2} \left|\left| \begin{pmatrix} f_k \\ f_k \end{pmatrix} - \begin{pmatrix} D_k \\ D_k \tilde{Q}_i \tilde{Q}_i^T \end{pmatrix} r \right|\right|_2^2 \\ &= \frac{1}{2} \left(||f_k - R_r \alpha_k||_2^2 + ||f_k - R_{\tilde{Q}_i \tilde{Q}_i^T r} \alpha_k||_2^2 \right) \\ &= \frac{1}{2} \alpha_k^T \left(R_r^T R_r + R_{\tilde{Q}_i \tilde{Q}_i^T r}^T R_{\tilde{Q}_i \tilde{Q}_i^T r} \right) \alpha_k \\ &\qquad - [(R_r^T + R_{\tilde{Q}_i \tilde{Q}_i^T r}^T) f_k]^T \alpha_k. \end{aligned}$$

Now, we sum over all sparse codes r obtained from the training set and incorporate objective (7) to obtain the overall program. Denote $\mathcal{R}_1 = \sum_r (R_r^T R_r + R_{\tilde{Q}_i \tilde{Q}_\ell^T r}^T R_{\tilde{Q}_\ell \tilde{Q}_\ell^T r}), \mathcal{R}_2 = \sum_r (R_r^T + R_{\tilde{Q}_\ell \tilde{Q}_\ell^T r}^T) f_k$ – an abuse of notation is committed where ℓ and f_k always denote, respectively, the associated class and graph signal (at channel k) of the sample sparse code r as we sum over the dataset. To incorporate (7) as a penalty, note that $\mathcal{Q}_\ell^p(g_{i,j}) = \frac{1}{L-1} ||\Omega_{/\ell} V_H \alpha_{i,j}||_p^p$, where V_H is the Vandermonde matrix assembled from the bin centers H. Our full dictionary update then reads as the unconstrained problem

$$\min_{\alpha_k} \frac{1}{2} \alpha_k^T \mathcal{R}_1 \alpha_k - \mathcal{R}_2^T \alpha_k + \frac{\eta}{L-1} \sum_{\ell=1}^{L} \sum_{j=(\ell-1)m}^{\ell m - 1} ||\Omega_{/\ell} V_H \alpha_{k,j}||_p^p. \tag{8}$$

[1] such density would only be attained at the limit of $|\mathcal{V}| \to \infty$, i.e. only the eigenvalues of an infinite graph would completely follow the distribution.

If $p = 2$, the problem is essentially an unconstrained quadratic program with an unorthodox L^2 regularization, which we can solve with the conjugate gradient method []. Setting $p = 1$ further increases the difficulty as we then deal with a non-smooth penalty. For a start, one might think to substitute $z := [\Omega_{/1} V_H, ..., \Omega_{/L} V_H]\alpha_k$ and solve

$$\begin{aligned} &\min_{\alpha_k, z} \frac{1}{2}\alpha_k \mathcal{R}_1 \alpha_k - \mathcal{R}_2^T \alpha_k + \frac{\eta}{L-1}||z||_1 \\ &\text{subject to } z - [\Omega_{/1} V_H, ..., \Omega_{/L} V_H]\alpha_k = 0 \end{aligned} \tag{9}$$

with the Alternating Direction Method of Multipliers (ADMM) []. For conciseness, however, we only consider the case $p = 2$.

B. Classification

Sparse coding coefficients r^* are computed by minimising $O(D, r)$ with the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [28]. The final task of classification given r^* can be made in a number of ways. Ref. [11] achieve best results with a classifier which picks the label

$$\arg\min_i \{||f - D\tilde{Q}_i \tilde{Q}_i^T r^*||_2^2 + \lambda ||\tilde{Q}_i^T r^*||_2^2\}.$$

Whether the sparse coding is done simultaneously over the whole dictionary (global classification) or individually over the sub-dictionaries (local classification) is a data-dependent decision. Here, we only consider classification with a multi-layer perceptron h_θ with weights and bias θ.

VI. Model optimization

A. Parameter updates

Dictionary kernel coefficients can be updated through mini-batched stochastic gradient descent on (6) with Adaptive Moment Estimation (Adam) [29]. Classifier parameters θ are optimised similarly, but the associated gradients now flow from the cross-entropy (CE) loss resulting from the classification error. One backward step consists of an update of θ proceeded by an update of $A = (\alpha_{1,1}, ..., \alpha_{C,M})$. The coefficients are not optimised to fit the cross-entropy loss so as to preserve interpretability of the kernels. Full-batch updates of A with (8) provide an interesting alternative if the spectral densities of each class have significantly differing concentrations.

B. Preprocessing transform

In cases where the dataset consists of relatively bare graphs requiring an informative nodal signal, a node embedding model with parameters φ is used to provide each graph with a high-dimensional signal as the first part of the forward step. Only then does the sparse coding and classification occur, after which the parameter update takes place, now including an update of the embedder model's parameters. The intermediate gradient $\partial_f(r^*)$ integral to the calculation of ∂_φ(CE loss) can be computed by implicit differentiation software or by autodifferentiation through the solver's unrolled updates. The parameters φ and θ share the same optimisation step, this last step preceding the optimisation step for A as mentioned previously.

VII. Experiments

A. Datasets

Six popular classification benchmarks, four of which (MUTAG, ENZYMES, PROTEIN, and NCI1) stem from bioinformatics, with the remaining two datasets (IMDB-B, COLLAB) characterising social network data. Two generated classes of attributed graphs – comprising as a whole one synthetic dataset – are devised to evaluate the effectiveness of overlap minimisation and full batch updating. One class consists exclusively of scale-free networks while the other class consists exclusively of Erdos-Reyni graphs. To each class is associated a set of 15 generative spectral kernels that is used to endow the class samples with a multi-channel signal.

B. Baselines

TODO

C. Performance and Visualisation

TODO

D. Ablation Study

TODO

VIII. Conclusion

TODO

References

[1] D. K. Hammond, P. Vandergheynst, and R. Gribonval. "Wavelets on Graphs via Spectral Graph Theory". *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129-150, 2011.

[2] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura and P. Vandergheynst, "Graph Signal Processing: Overview, Challenges, and Applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808-828, May 2018.

[3] A. Gavili and X. -P. Zhang, "On the Shift Operator, Graph Frequency, and Optimal Filtering in Graph Signal Processing," *IEEE Transactions on Signal Processing*, vol. 65, no. 23, pp. 6303-6318, 1 Dec.1, 2017.

[4] Y. Tanaka, Y. C. Eldar, A. Ortega and G. Cheung, "Sampling Signals on Graphs: From Theory to Applications," *IEEE Signal Processing Magazine*, vol. 37, no. 6, pp. 14-30, Nov. 2020.

[5] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4-24, Jan. 2021.

[6] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *ICLR,* 2017, pp. 1-14.

[7] L, Chuang, Y. Zhan, J. Wu, C. Li, B. Du, W. Hu, T. Liu, and D. Tao, "Graph pooling for graph neural networks: Progress, challenges, and opportunities." in *IJCAI*, 2023, pp.1-11.

[8] M. M. Bronstein, J. Bruna, T. Cohen, P. Veličković, "Geometric Deep Learning Models" in *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*, 1st ed.(TODO: May 4, 2021 by arxiv), pp. 77-80.

[9] Z. Zhang et al., "Hierarchical Multi-View Graph Pooling With Structure Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 545-559, 1 Jan. 2023.

[10] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210-227, Feb. 2009.

[11] S. Kong and D. Wang, "A Dictionary Learning Approach for Classification: Separating the Particularity and the Commonality," in *Computer Vision - ECCV 2012*, Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2012, pp. 186-199.

[12] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, "Discriminative learned dictionaries for local image analysis," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1-8.
[13] K. Skretting, and J. H. Husøy. "Texture classification using sparse frame-based representations." EURASIP Journal on Advances in Signal Processing 2006, no. 1
[14] R. Grosse, R. Raina, H. Kwong, and A. Ng, "Shift-Invariant Sparse Coding for Audio Classification," in *UAI,* 2007, pp. 149-158.
[15] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems 21*, 2008.
[16] Z. Jiang, Z. Lin and L. S. Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651-2664, Nov. 2013.
[17] M. Yang, L. Zhang, X. Feng and D. Zhang, "Fisher Discrimination Dictionary Learning for sparse representation," in *2011 International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 543-550.
[18] K. Huang and S. Aviyente, "Sparse Representation for Signal Classification," in *Neural Information Processing Systems*, 2006.
[19] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *CVPR*, 2010.
[20] X. Zhang, X. Dong and P. Frossard, "Learning of structured graph dictionaries," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 3373-3376.
[21] 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, Aug.1, 2014.
[22] O. Christensen, "A short introduction to frames, Gabor systems, and wavelet systems," *Azerbaijan Journal of Mathematics*, pp. 25-39, 2014.
[23] C. Garcia-Cardona and B. Wohlberg, "Convolutional Dictionary Learning: A Comparative Review and New Algorithms," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 366-381, Sept. 2018.
[24] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *Applied and Computational Harmonic Analysis*, vol. 40, no. 2, pp. 260-291, 2016.
[25] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation Learning on Graphs: Methods and Applications," *IEEE Data Engineering Bulletin*, vol. 40, pp. 52-74, 2017.
[26] S. Kim, R. Pasupathy, and S. G. Henderson, "A Guide to Sample Average Approximation," in *Handbook of Simulation Optimization*, M. C. Fu. Editor, Ed. 127, Springer, 2015, pp. 207-243.
[27] N. S. Chatterji and P. L. Barnett, "Alternating minimization for dictionary learning with random initialization," in *31st Conference on Neural Information Processing Systems (NIPS 2017*, 2017.
[28] A. Beck and M. Teboulle, "A fast Iterative Shrinkage-Thresholding Algorithm with application to wavelet-based image deblurring," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 693-696.
[29] D. P. Kingma and J. Ba, "Adam: a Method for Stochastic Optimization," in *3rd International Conference on Learning Representations (ICLR 2015)*, 2015.