Learning Nonseparable Sparse Regularizers via Multivariate Activation Functions (Appendix)

Anonymous Authors[1]

A. Conditions for Equation (17)

Associated with the explicit form (16) (of the manuscript) of

$$
\begin{aligned}
&g(x)\\
&=\begin{cases}
\left(\frac{1}{2\eta_2}-\frac{1}{2}\right)x^2 & \\
+\left(\delta_2-\frac{\eta_1(\delta_2-\delta_1)}{\eta_2}\right)x & \\
+\frac{\eta_1(\eta_1-\eta_2)}{2\eta_2}\left(\delta_2-\delta_1\right)^2, & x\geq\eta_1\left(\delta_2-\delta_1\right),\\
\left(\frac{1}{2\eta_1}-\frac{1}{2}\right)x^2+\delta_1 x, & 0\leq x<\eta_1\left(\delta_2-\delta_1\right),\\
g(-x), & x<0,
\end{cases}
\end{aligned}
\tag{1}
$$

to be learned, more details for the deduction of conditions (17) (of the manuscript) for the learned parameters are provided. We consider the following two cases:

Case 1 : Suppose $0\leq x<\eta_1\left(\delta_2-\delta_1\right)$, it is evident that we have $g(x)=\left(\frac{1}{2y_1}-\frac{1}{2}\right)x^2+\delta_1 x\geq 0$ when $\eta_1\in(0,1]$. If $\eta_1\in(1,+\infty)$, the condition for $g(x)\geq 0$ becomes

$$\left(\frac{1}{2\eta_1}-\frac{1}{2}\right)\left(\eta_1\left(\delta_2-\delta_1\right)\right)^2+\delta_1\left(\eta_1\left(\delta_2-\delta_1\right)\right)\geq 0, \tag{2}$$

this is,

$$\left(\eta_1-1\right)\delta_2-\left(\eta_1+1\right)\delta_1\leq 0. \tag{3}$$

Case 2: Suppose $x\geq\eta_1\left(\delta_2-\delta_1\right)$, because $\frac{1}{2\eta_2}-\frac{1}{2}<0$ when $\eta_2\in(1,+\infty)$, it can be verified that $g(x)<0$ when x is large enough. Therefore we only consider $\eta_2\in(0,1]$. It is required that $g(x)$ is non-decreasing for $x\in[\eta_1\left(\delta_2-\delta_1\right),+\infty)$, leading to $\nabla g(x)\geq 0$. According to the expression of $g(x)$ given in (16) (of the manuscript), we obtain

$$\nabla g(x)=\left(\frac{1}{\eta_2}-1\right)x+\left(\delta_2-\frac{\eta_1\left(\delta_2-\delta_1\right)}{\eta_2}\right). \tag{4}$$

Together with $\eta_2\in(0,1]$, we know that $\nabla g(x)$ is also non-decreasing when $x\geq\eta_1\left(\delta_2-\delta_1\right)$. Therefore, $\nabla g(x)\geq 0$ for $x\in[\eta_1\left(\delta_2-\delta_1\right),+\infty)$ is equivalent to $\nabla g(x)|_{x=\eta_1(\delta_2-\delta_1)}\geq 0$, resulting in

$$
\begin{aligned}
&\left(\frac{1}{\eta_2}-1\right)\eta_1\left(\delta_2-\delta_1\right)+\left(\delta_2-\frac{\eta_1\left(\delta_2-\delta_1\right)}{\eta_2}\right)\\
&=-\,\eta_1\left(\delta_2-\delta_1\right)+\delta_2\geq 0,
\end{aligned}
\tag{5}
$$

which indicates $\delta_1\geq\frac{\eta_1-1}{\eta_1}\delta_2$. Because $g(x)$ is non-decreasing and $g\left(\eta_1\left(\delta_2-\delta_1\right)\right)\geq 0$ when $\delta_1\geq\frac{\eta_1-1}{\eta_1}\delta_2$, we have $g(x)\geq 0$ in this case. Simultaneously,

$$\delta_1\geq\frac{\eta_1-1}{\eta_1}\delta_2>\frac{\eta_1-1}{\eta_1+1}\delta_2, \tag{6}$$

which also guarantees that Inequality (3) holds. In summary, combing $\eta_1,\eta_2>0,\delta_2\geq\delta_1>0$ from the aforementioned analyses, the conditions for making $g(x)$ nonnegative become

$$
\begin{aligned}
&\eta_1>0,1\geq\eta_2>0,\\
&\delta_2\geq\delta_1\geq\max\left\{0,\tfrac{\eta_1-1}{\eta_1}\delta_2\right\}.
\end{aligned}
\tag{7}
$$

■

B. The Proofs of (19) and (20)

In this section, we prove that the projection $\operatorname{Proj}\left(\delta_1,\delta_2\right)$ of $\left(\delta_1,\delta_2\right)$ onto $\mathcal{S}_\delta$ is (19) or (20) (of the manuscript). When $0<\eta_1\leq 1$, as shown in Figure 1, we consider the following four cases:

Case 1: When $\left(\delta_1,\delta_2\right)$ is in region A, namely $0\leq\delta_1\leq\delta_2$, as shown in the blue area of Figure 1, the projection $\operatorname{Proj}\left(\delta_1,\delta_2\right)$ is $\left(\delta_1,\delta_2\right)$.

Case 2: When $\left(\delta_1,\delta_2\right)$ is in region B, namely $\delta_2>0>\delta_1$, the projection $\operatorname{Proj}\left(\delta_1,\delta_2\right)$ is on the δ_2-axis, i.e., $\left(0,\delta_2\right)$.

Case 3: When $\left(\delta_1,\delta_2\right)$ is in region C, namely $\delta_2\leq\min\left\{0,-\delta_1\right\}$, the projection $\operatorname{Proj}\left(\delta_1,\delta_2\right)$ is at the origin $(0,0)$.

Case 4: When $\left(\delta_1,\delta_2\right)$ is in region D, namely $\delta_1\geq\left|\delta_2\right|$, the projection $\operatorname{Proj}\left(\delta_1,\delta_2\right)$ is on the boundary between region A and region D, i.e., $\left(\frac{\delta_1+\delta_2}{2},\frac{\delta_1+\delta_2}{2}\right)$.

To sum up, when $0<\eta_1\leq 1$, the projection $\operatorname{Proj}\left(\delta_1,\delta_2\right)$

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

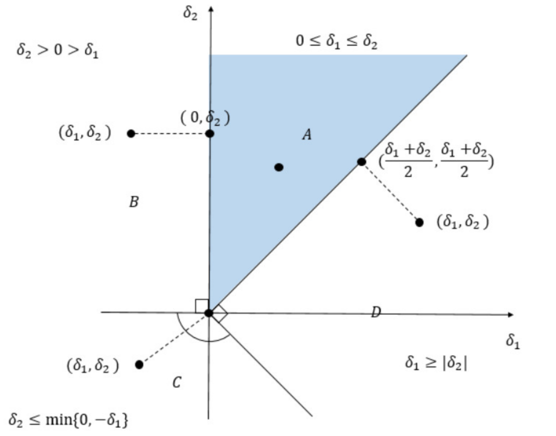

δ_2
$\delta_2 > 0 > \delta_1$
$0 \leq \delta_1 \leq \delta_2$
$(0, \delta_2)$
(δ_1, δ_2)
A
$(\frac{\delta_1+\delta_2}{2}, \frac{\delta_1+\delta_2}{2})$
B
(δ_1, δ_2)
D
δ_1
(δ_1, δ_2)
$\delta_1 \geq |\delta_2|$
C
$\delta_2 \leq \min\{0, -\delta_1\}$

Figure 1. The projection $\operatorname{Proj}(\delta_1, \delta_2)$ of (δ_1, δ_2) onto $\mathcal{S}_\delta$ $(0 < \eta_1 \leq 1)$

of (δ_1, δ_2) onto $\mathcal{S}_\delta$ is

$$\begin{aligned} &\operatorname{Proj}(\delta_1, \delta_2) \\ &= \begin{cases} (\delta_1, \delta_2), & \delta_1 \geq 0, \delta_2 \geq 0, \delta_1 \leq \delta_2, \\ (0, \delta_2), & \delta_1 < 0, \delta_2 > 0, \\ (0, 0), & \delta_2 \leq \min\{0, -\delta_1\}, \\ \left(\frac{\delta_1+\delta_2}{2}, \frac{\delta_1+\delta_2}{2}\right), & \delta_1 \geq |\delta_2|. \end{cases} \end{aligned} \tag{8}$$

In addition, when $\eta_1 > 1$, in a similar way, we can get the projection $\operatorname{Proj}(\delta_1, \delta_2)$ as

$$\begin{aligned} &\operatorname{Proj}(\delta_1, \delta_2) \\ &= \begin{cases} (\delta_1, \delta_2), & \delta_2 \geq 0, \frac{\eta_1 - 1}{\eta_1}\delta_2 \leq \delta_1 \leq \delta_2, \\ (\rho_1\delta_1 + \rho_2\delta_2, \rho_2\delta_1 & \\ \quad +\rho_3\delta_2), & \frac{\eta_1}{1-\eta_1}\delta_2 < \delta_1 < \frac{\eta_1-1}{\eta_1}\delta_2, \\ (0, 0), & \delta_2 \geq 0, \delta_1 \leq \frac{\eta_1}{1-\eta_1}\delta_2, \\ (0, 0), & \delta_2 \leq \min\{0, -\delta_1\}, \\ \left(\frac{\delta_1+\delta_2}{2}, \frac{\delta_1+\delta_2}{2}\right), & \delta_1 \geq |\delta_2|, \end{cases} \end{aligned} \tag{9}$$

where the parameter $\{\rho_1, \rho_2, \rho_3\}$ is given as $\rho_1 = \frac{(\eta_1-1)^2}{\eta_1^2+(\eta_1-1)^2}$, $\rho_2 = \frac{\eta_1(\eta_1-1)}{\eta_1^2+(\eta_1-1)^2}$ and $\rho_3 = \frac{\eta_1^2}{\eta_1^2+(\eta_1-1)^2}$.

C. Experiments of Multi-view Clustering

We conduct comprehensive experiments on publicly available real-world datasets to validate the superiority of the learned sparse regularizer by MAF-SRL in terms of multi-view clustering.

Given multi-view data $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^v$ with $\mathbf{x}_i \in \mathbb{R}^{n \times d_i}$ where n and v are the sample and view numbers, and d_i is the feature number of the i-th view data. Consequently, the multi-view clustering task is to learn a cluster indicator $\mathbf{y} \in \{0, 1\}^n$ from the given multi-view data with certain criterion loss $(\{\mathbf{x}_i\}_{i=1}^v; \mathbf{y})$. Considering that different views may come with varying dimensions, we attempt to learn an optimal affinity matrix from the evaluated multi-view similarity matrices $\mathcal{W} = \{\mathbf{W}_i\}_{i=1}^v$ of $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^v$. In order to verify the superiority of the proposed learnable sparse regularizer method, we formulate the multi-view clustering task as the following simple form

$$\begin{aligned} \arg\min\nolimits_{\boldsymbol{\alpha}, \mathbf{W}} \tfrac{1}{2} \left\| \mathbf{W} - \textstyle\sum_{j=1}^v \alpha_j \mathbf{W}_j \right\|_F^2 + g(\mathbf{W}), \\ \text{subject to} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}, \boldsymbol{\alpha}^T \mathbf{1} = 1, \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha} = [\alpha_1; \cdots; \alpha_v] \in \mathbb{R}^v$ is a v-dimensional column vector representing the weights of all views, and $g(\cdot)$ is a sparse regularizer yet to be learned. The fused affinity matrix of multi-view data is represented as a convex hull of all views, and the representation coefficients are learned adaptively from the optimization objective. Since the view number v tends to be small, a separate algorithm is developed to compute the optimal value of $\boldsymbol{\alpha}$. And adaptive weights can be optimized by the ADMM algorithm. Particularly, suppose that $\operatorname{vec}(\cdot)$ is the matrix vectorization operator, then the optimization subproblem with respect to α is written as

$$\begin{aligned} \min\nolimits_\alpha \tfrac{1}{2} \left\| [\operatorname{vec}(\mathbf{W}_1), \cdots, \operatorname{vec}(\mathbf{W}_v)] \boldsymbol{\alpha} - \operatorname{vec}(\mathbf{W}) \right\|_F^2 \\ \text{subject to} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}, \boldsymbol{\alpha}^T \mathbf{1} = 1. \end{aligned} \tag{11}$$

While keeping the weighted vector $\boldsymbol{\alpha}$, we compute the optimal solution $\mathbf{W} = \operatorname{Prox}_g^\sigma \left(\sum_{j=1} \alpha_j \mathbf{W}_j \right)$. Because $\phi(\mathbf{W}) = \frac{1}{2} \left\| \mathbf{W} - \sum_{j=1}^v \alpha_j \mathbf{W}_j \right\|_F^2$ is differentiable, we can apply the proposed MAF-SRL framework to learn an optimal data-driven sparse regularizer $g(\mathbf{W})$.

C.1 Datasets

In this subsection, eight publicly available datasets are used to validate the effectiveness of the proposed method MAF-SRL. These datasets are derived from real-world image applications, ranging from images to videos. Several sample images are randomly collected from the test datasets, as demonstrated in Figure 2. It is suggested from the figure that different images may be captured with varied viewing angles, illumination colors and resolution variations, which motivates us to extract multi-view low-level features using feature descriptors for varying datasets. Here we provide more details for the feature extractors of these test datasets.

ALOI[1]: A collection of object images were taken under varied light conditions and rotation angles. Four commonly used features include 64-dimension RGB color histograms, 64 dimension HSV color histograms, 77-dimension color similarities and 13-dimension Haralick features.

[1] https://elki-project.github.io/datasets/multi view

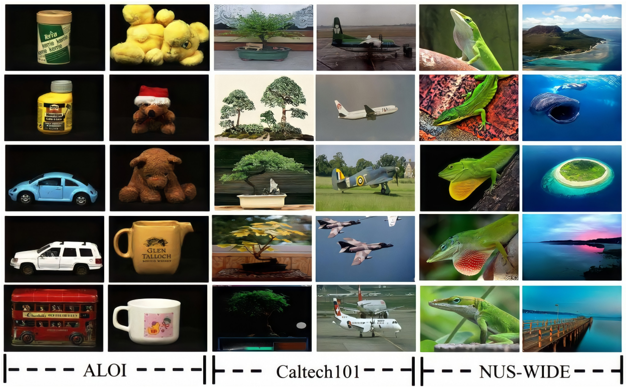


Figure 2. Several sample images from the test image datasets.

Caltech101-7/Caltech101-20[2]: Caltech101 is a popular object recognition dataset with 101 classes of images. We follow previous work and select the widely used subsets Caltech101-7 and Caltech101-20. Six extracted features are used: 48-dimension Gabor, 40-dimension wavelet moments (WM), 254-dimension CENTRIST, 1,984-dimension histogram of oriented gradients (HOG), 512-dimension GIST and 928-dimension LBP features.

MNIST[3]: It is a well known dataset of handwritten digits. Three types of features are extracted from all test images: 30-dimension IsoProjection, 9-dimension Linear Discriminant Analysis (LDA) and 9-dimension Neighborhood Preserving Embedding (NPE) features.

NUS-WIDE[4]: As a web image dataset for object recognition, we select eight classes and six available feature sets: 64 dimension color histogram, 225-dimension block-wise color moments, 144-dimension color correlogram, 73-dimension edge direction histogram, 128-dimension wavelet texture and 500-dimension bag of words from SIFT descriptors.

MSRC-v1[5]: It is an image dataset with 8 object classes and each class has 30 images. We select 7 classes composed of tree, building, airplane, cow, face, car and bicycle. Five visual feature sources are extracted from each image: 24-dimension color moment, 576-dimension HOG, 512-dimension GIST, 256-dimension local binary pattern and 256-dimension CENTRIST features.

ORL[6]: This database contains ten different face images, each of 40 subjects, which were taken at various times, differing in the lighting and facial expressions.

Youtube[7]: It is a video dataset with 2,000 instances in 10 topics, along with six views from both visual features and audio features, including 2,000-dimension cuboids histogram, 1,024 dimension hist motion estimate, 64-dimension HOG features, 512-dimension MFCC features, 64-dimension volume streams, and 647-dimension spectrogram streams.

C.2 Performance Evaluation

For clustering tasks, three well-known evaluation metrics including clustering accuracy (ACC), normalized mutual information (NMI) and adjusted rand index (ARI) are applied to the comparative experiments. Given any sample $\mathbf{x}_i$ the predicted clustering label and the real label are denoted by p_i and q_i, respectively. The ACC is defined as

$$\mathrm{ACC}=\frac{\sum_{i=1}^{n}\delta\left(q_i,\operatorname{map}\left(p_i\right)\right)}{n}, \tag{12}$$

where $\delta(a,b)=1$ if $a=b$, and $\delta(a,b)=0$ otherwise. Here, $\operatorname{map}(\cdot)$ is the best permutation mapping that matches the predicted clustering labels to the ground truths. Denote the predictive clustering result as $\widehat{\mathbf{C}}=\left\{\widehat{\mathbf{C}}_i\right\}_{i=1}^{\widehat{c}}$ and the ground truth as $\mathbf{C}=\{\mathbf{C}_j\}_{j=1}^{c}$, NMI is calculated by

$$\mathrm{NMI}=\frac{\sum_{i=1}^{\widehat{c}}\sum_{j=1}^{c}\left|\widehat{\mathbf{C}}_i\cap\mathbf{C}_j\right|\log\frac{n\left|\widehat{\mathbf{C}}_i\cap\mathbf{C}_j\right|}{|\mathbf{C}_i||\mathbf{C}_j|}}{\sqrt{\left(\sum_{i=1}^{\widehat{c}}\left|\widehat{\mathbf{C}}_i\right|\log\frac{|\widehat{\mathbf{C}}_i|}{n}\right)\left(\sum_{j=1}^{c}|\mathbf{C}_j|\log\frac{|\mathbf{C}_j|}{n}\right)}}. \tag{13}$$

And ARI characterizes the agreement between two partitions $\mathbf{C}$ and $\widehat{\mathbf{C}}$, defined as

$$\mathrm{ARI}=\frac{\sum_{ij}N_{ij}-\left[\sum_i A_i\sum_j B_j\right]/\begin{pmatrix}n\\2\end{pmatrix}}{\frac{1}{2}\left[\sum_i A_i+\sum_j B_j\right]-\left[\sum_i A_i\sum_j B_j\right]/\begin{pmatrix}n\\2\end{pmatrix}}, \tag{14}$$

where $N_{ij}=\begin{pmatrix}n_{ij}\\2\end{pmatrix}$, $A_i=\begin{pmatrix}a_i\\2\end{pmatrix}$, $B_j=\begin{pmatrix}b_j\\2\end{pmatrix}$, $[n_{ij}]=\left|\widehat{\mathbf{C}}_i\cap\mathbf{C}_j\right|$, $a_i=\left|\widehat{\mathbf{C}}_i\right|$ and $b_j=|\mathbf{C}_j|$. The higher values of all these metrics indicate the better performance. All experiments are repeated 20 times, and we report the means as the final results.

C.3 Multi-View Clustering

As for the proposed method MAF-SRL, the activation function defined in (9) (of the manuscript), i.e., $\xi(\mathbf{x})=\mathbf{A}^{-T}\left[\widehat{\xi}\left((\mathbf{A}\operatorname{diag}(\boldsymbol{q}))^{-1}\mathbf{x}\right)-\mathbf{b}\right]$, is employed. An initialization for the parameterized activation function is tuned as $\eta_1=\eta_2=1.0$, $\delta_1=1.0$ and $\delta_2=2.0$.

[2] http://www.vision.caltech.edu/Image Datasets/Caltech101/

[3] http://yann.lecun.com/exdb/mnist/

[4] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUSWIDE.html

[5] http://riemenschneider.hayko.at/vision/dataset/task.php?did=35

[6] http://cam-orl.co.uk/facedatabase.html

[7] http://archive.ics.uci.edu/ml/datasets

Table 1. Clustering accuracy with different sparse regularizers, where the best performance is highlighted in bold.

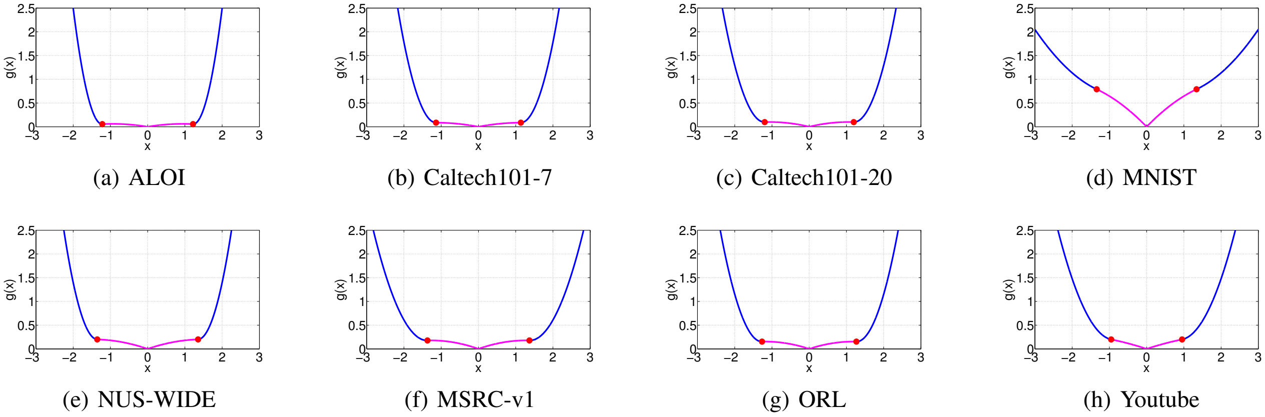
Dataset	Metrics	ℓ_1	ℓ_{1-2}	SGL	CGES	SCAD	capped-ℓ_1	LSP	MCP	MAF-SRL (ours)
ALOI	ACC	0.6538	0.7146	0.6637	0.7129	0.6329	0.7792	0.7349	0.7839	**0.8374**
	NMI	0.7473	0.7639	0.6993	0.7742	0.6395	0.7983	0.7539	0.7439	**0.8849**
	ARI	0.6521	0.6029	0.5983	0.6849	0.6175	0.6844	0.6833	0.5948	**0.7219**
Caltech101-7	ACC	0.7129	0.6674	0.8129	0.7749	0.8022	0.7375	0.8355	0.7983	**0.8935**
	NMI	0.6639	0.7174	0.6893	0.7112	0.7329	0.7121	0.6899	0.7019	**0.7495**
	ARI	0.5948	0.6849	0.5584	0.6439	0.5992	0.6217	0.6549	0.6493	**0.7042**
Caltech101-20	ACC	0.7753	0.6139	0.7399	0.6539	0.6926	0.7893	0.7837	0.8127	**0.8539**
	NMI	0.6783	0.6648	0.7129	0.7583	0.6929	0.6327	0.6349	0.6649	**0.8003**
	ARI	0.6938	0.7093	0.6928	0.5947	0.6548	0.7326	0.7133	0.7247	**0.7749**
MNIST	ACC	0.8031	0.7749	0.7388	0.6928	0.7449	0.8022	0.7837	0.7762	**0.8636**
	NMI	0.7233	0.6649	0.6538	0.6644	0.6291	0.7129	0.7459	0.7837	**0.8329**
	ARI	0.7749	0.6938	0.6554	0.7034	0.7592	0.7783	0.6846	0.7206	**0.8022**
NUS-WIDE	ACC	0.5053	0.4927	0.4872	0.4551	0.4029	0.3993	0.4892	0.4463	**0.5339**
	NMI	0.1948	0.2984	0.3336	0.2841	0.3083	0.2943	0.2988	0.2875	**0.3992**
	ARI	0.3294	0.3847	0.2998	0.3736	0.2988	0.4539	0.3352	0.4029	**0.4873**
MSRC-v1	ACC	0.8392	0.7539	0.7733	0.8024	0.7938	0.8147	0.8473	0.7939	**0.8893**
	NMI	0.6547	0.7749	0.7328	0.7459	0.7201	0.6993	0.7023	0.6994	**0.7938**
	ARI	0.6839	0.7055	0.7649	0.7351	0.6694	0.7639	0.7627	0.6891	**0.8036**
ORL	ACC	0.8732	0.7793	0.8265	0.8004	0.8501	0.8837	0.7322	0.7837	**0.9038**
	NMI	0.7935	0.8118	0.7894	0.8227	0.8092	0.7684	0.8106	0.7787	**0.8547**
	ARI	0.6839	0.7128	0.8005	0.7739	0.6845	0.7493	0.6949	0.7239	**0.8458**
Youtube	ACC	0.4297	0.4471	0.5076	0.5309	0.4727	0.5574	0.5157	0.6029	**0.6249**
	NMI	0.4893	0.5092	0.4474	0.3981	0.4983	0.5427	0.4971	0.4925	**0.5882**
	ARI	0.1939	0.3742	0.2947	0.3903	0.3131	0.2992	0.3832	0.3981	**0.4585**

Figure 3 shows the learned univariate function $g(x)$ by the activation function $\xi(\mathbf{x})$ of MAF-SRL on all test datasets for clustering tasks. The learned parameters $\{\eta_1, \eta_2, \delta_1, \delta_2\}$ differ in varied datasets, as a result of learning sparse regularizers in a data-driven manner. All learned parameters of activation functions obey (17) (of the manuscript).

Table 1 records the clustering accuracy on all test datasets with different sparse regularizers for ACC, NMI and ARI. From the experimental results, we have the observation that the proposed MAF-SRL achieves best performance than all compared manually designed sparse regularizers. Table 2 further provides the sparsity of the learned sparse regularizers, where sparsity is defined by the ratio of nonzero weights in the backbone network.

The performance of MAF-SRL for clustering tasks is reported in Figure 4 with ACC, NMI and ARI as the number of layers varies. For all figures, the layer number ranges in $\{2, 4, \cdots, 30\}$, and the learning rate lr is fixed as 0.15. Consistent with previous experiments on classification tasks, the accuracy roughly increases with more layers and becomes stable when the layer number $N > 16$.

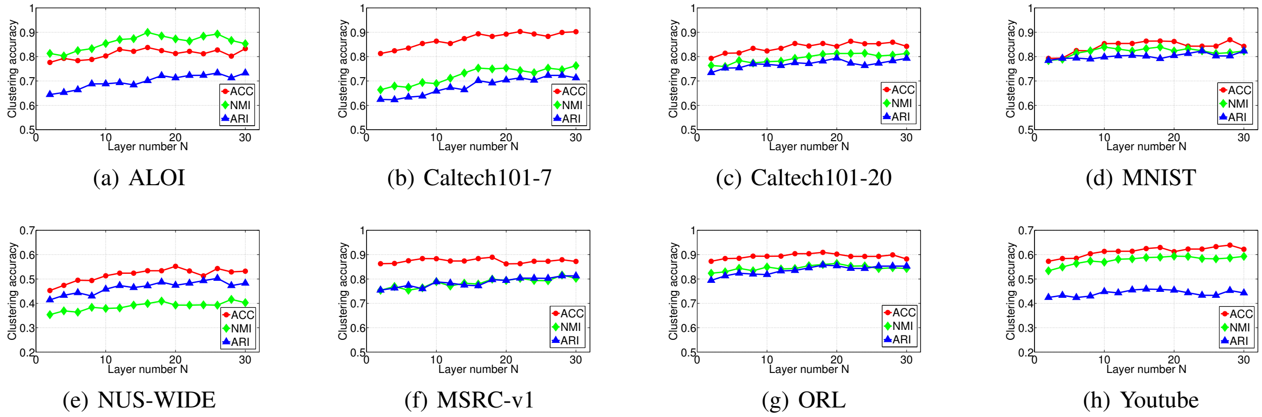
We can find that MAF-SRL achieves the best performance with the best sparse outputs (lowest percentage of nonzero weights). These observations indicate that our proposed multivariate activation functions succeed in learning a data-driven sparse representation of similarity matrices, and thus such learned sparse regularizers are more robust when applied to various datasets. MAF-SRL can learn a data-driven sparse regularizer, which naturally exhibits strong generalization capability in practical applications.



(a) ALOI (b) Caltech101-7 (c) Caltech101-20 (d) MNIST

(e) NUS-WIDE (f) MSRC-v1 (g) ORL (h) Youtube

Figure 3. The learned univariate function $g(x)$ on different datasets for multi-view clustering. Its associated parameters are as follows: (a) $\eta_1 = 1.16, \eta_2 = 0.11, \delta_1 = 0.13, \delta_2 = 1.18$. (b) $\eta_1 = 1.15, \eta_2 = 0.18, \delta_1 = 0.15, \delta_2 = 1.14$. (c) $\eta_1 = 1.22, \eta_2 = 0.21, \delta_1 = 0.19, \delta_2 = 1.17$. (d) $\eta_1 = 1.49, \eta_2 = 0.68, \delta_1 = 0.81, \delta_2 = 1.71$. (e) $\eta_1 = 1.22, \eta_2 = 0.15, \delta_1 = 0.27, \delta_2 = 1.38$. (f) $\eta_1 = 1.28, \eta_2 = 0.31, \delta_1 = 0.28, \delta_2 = 1.35$. (g) $\eta_1 = 1.28, \eta_2 = 0.21, \delta_1 = 0.26, \delta_2 = 1.25$. (h) $\eta_1 = 1.12, \eta_2 = 0.33, \delta_1 = 0.26, \delta_2 = 1.11$.



(a) ALOI (b) Caltech101-7 (c) Caltech101-20 (d) MNIST

(e) NUS-WIDE (f) MSRC-v1 (g) ORL (h) Youtube

Figure 4. The relations among clustering accuracy (ACC, ARI and NMI) and layer number in $\{2, 4, \cdots, 30\}$ of the proposed method MAF-SRL.

Table 2. The weight sparsity of different methods on the datasets for multi-view clustering.

Dataset	ℓ_1	ℓ_{1-2}	SGL	CGES	SCAD	capped-ℓ_1	LSP	MCP	MAF-SRL (ours)
ALOI	0.2849	0.2283	0.2937	0.2827	0.2948	0.2301	0.2529	0.2729	**0.1638**
Caltech101-7	0.2039	0.2312	0.2574	0.2739	0.2029	0.2938	0.2993	0.2395	**0.1801**
Caltech101-20	0.2648	0.2947	0.2357	0.3003	0.2995	0.3021	0.2854	0.2836	**0.2011**
MNIST	0.2029	0.2227	0.2518	0.2922	0.1988	0.2022	0.2837	0.1962	**0.1643**
NUS-WIDE	0.2936	0.3315	0.3079	0.3128	0.2975	0.3287	0.3762	0.3529	**0.2584**
MSRC-v1	0.2531	0.3128	0.2483	0.2839	0.2617	0.2491	0.2148	0.2455	**0.1938**
ORL	0.1321	0.1732	0.1206	0.1995	0.2012	0.1883	0.1381	0.1937	**0.0929**
Youtube	0.2947	0.2348	0.2865	0.2917	0.2649	0.2833	0.2846	0.2753	**0.1975**