Authors' Responses to the Review of Manuscript CYB-E-2019-03-0631 "Scaled Simplex Representation for Subspace Clustering"

Jun Xu, Mengyang Yu, Ling Shao, *Senior Member*, IEEE, Wangmeng Zuo, *Senior Member*, IEEE,
Deyu Meng, Lei Zhang, *Fellow*, IEEE, David Zhang, *Fellow*, IEEE

Dear Editors and Reviewers,

We are appreciated to know that our manuscript was rated as potentially acceptable for publication in this journal, subject to a thorough revision. Thanks to the editors for providing us the opportunity to revise our manuscript. We sincerely thank the four reviewers and the associate editor for their enthusiasm and thoughtful feedback. We have carefully revised our manuscript, taking all the valuable comments and suggestions into consideration. Our point-by-point responses to all questions are given below.

Best Regards!

The Authors.

Responses to Reviewer 1

Recommendation: Accept

> ***Comment 1.1**: Minor comments on Page 2 and Page 6.*

Authors' response: Thank you very much for your wholeheartedly positive comments on our manuscript. We feel sorry to have typos in our submission, and have corrected them (**Page 2**, right column; **Page 6**, right column; colored by red) in the revision.

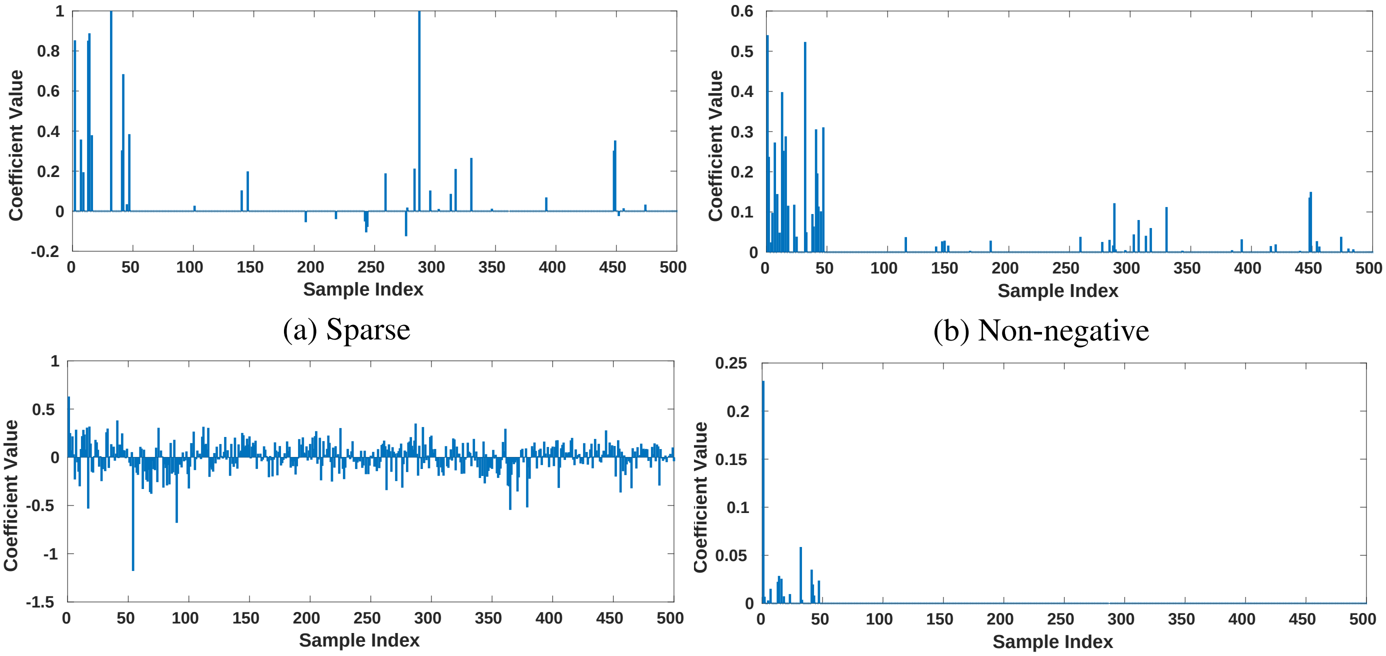
Responses to Reviewer 2

Recommendation: Prepare A Major Revision

> ***Comment 2.1**: An in-depth explanation of why affine scalar s works? Maybe numeric reason? what if scale it back to 1 after optimization?*

Authors' response: Thanks for the suggestion.

To illustrate the advantages of the proposed scaled simplex representation, in Figure 1 we show the comparison of coefficient vectors solved by different representation schemes on the handwritten digit images



(a) Sparse

(b) Non-negative

(c) Affine

(d) Scaled Simplex

Figure 1: Comparison of coefficient vectors by different representation schemes on the handwritten digit images from MNIST [1]. A digit sample "0" is used to compute the vector over 500 samples of digits $\{0, 1, ..., 9\}$ (50 for each). (a) The vector solved by a sparse model (e.g., LASSO [2]) are confusing. (b) The vector solved by the least square regression (LSR) model with non-negative constraint are noisy. (c) The coefficients solved by LSR with affine constraint are chaotic. (d) The LSR with the proposed scaled simplex representation (SSR) can obtain physically more meaningful coefficients vector.

from MNIST [1]. A digit sample "0" is used to compute the vector over 500 samples of digits $\{0, 1, ..., 9\}$ (50 for each). We observe that: The vector solved by a sparse model (e.g., LASSO [2]) are confusing, the coefficients over the samples of other digits are also non-zero (Figure 1 (a)). The vector solved by the least square regression (LSR) model with non-negative constraint are noisy (Figure 1 (b)). The coefficients solved by LSR with affine constraint are densely distributed over all samples (Figure 1 (c)). The LSR with the proposed scaled simplex representation (SSR) can obtain physically more meaningful coefficients vector (Figure 1 (d)). These discussions are added to the "Introduction" section (**Page 2**, left column, colored by red).

We also explained empirically in Section III-D why suppressing the scalar s works. Besides these explanation, we also demonstrate the effectiveness of the advantages of scaled simplex (with scalar $s < 1$) over the standard simplex (with scalar $s = 1$) in Figures 7, 8, and 9 of the revised manuscript. As can be seen, the SSRSC with $s < 1$ always achieves better performance over that with the constraint of $s = 1$.

The key is that when the coefficient vector is constrained to be non-negative and summed up to a scalar $s < 1$ during optimization, it (so the affinity matrix) enjoys more discriminative property when we solve the SSRSC model. Theoretically, scaling the sum of coefficient vector back to 1 does not influence the discriminative property obtained during the optimization with scalar $s < 1$.

> ***Comment 2.2**: How do these constraints work with other models, such as SSC, LRR, rather than just LSR.*

Authors' response: Our goal in this work is to design a novel while simple subspace clustering model, based the proposed scaled simplex representation (SSR). However, adding the SSR constraints into the SSC or LRR models is quite challenging, since they already have meaningful constraints. For example, the SSC is usually formulated as:

$$\min_{\boldsymbol{C}} \|\boldsymbol{C}\|_1 \quad \text{s.t.} \quad \boldsymbol{X} = \boldsymbol{X}\boldsymbol{C}, \mathbf{1}^\top \boldsymbol{C} = \mathbf{1}^\top, \operatorname{diag}(\boldsymbol{C}) = \mathbf{0}. \tag{1}$$

The affine constraint "$\mathbf{1}^\top \boldsymbol{C} = \mathbf{1}^\top$" and the proposed scaled affine constraint '$\mathbf{1}^\top \boldsymbol{C} = s\mathbf{1}^\top$" cannot be valid simultaneously. Besides, if we add the non-negative constraint to the SSC, its objective function $\|\boldsymbol{C}\|_1$ in Eqn. (1) is constant, i.e., each column of $\|\boldsymbol{C}\|_1$ becomes 1. Therefore, the SSR constraint cannot be directly added to the SSC.

Similarly, the LRR model is usually formulated as:

$$\min_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_{2,1} + \lambda \|\boldsymbol{C}\|_*. \tag{2}$$

The LRR with additional simplex constraints is very complex to be solved, when compared to the LSR version. This is especially true when considering that solving LRR usually needs SVD decomposition, which is computational expensive.

Therefore, we employed LSR due to its simplicity over SSC and LRR: LSR has no constraint and can be solved in closed-form [3].

> ***Comment 2.3**: How does the self-affinity influence the model? Maybe considering evaluating it in the ablation study?*

Authors' response: Thanks for this helpful suggestion. To study how the self-affinity influences the SSRSC model, we added the suggested self-affine constraint of "$\operatorname{diag}(\boldsymbol{C}) = \mathbf{0}$" to the proposed SSRSC. The novel variant is called "SSRSC-diag" by us and formulated as

$$\min_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda \|\boldsymbol{C}\|_F^2, \text{ s.t. } \boldsymbol{C} \geq 0, \mathbf{1}^\top \boldsymbol{C} = s\mathbf{1}^\top, \operatorname{diag}(\boldsymbol{C}) = \mathbf{0}. \tag{3}$$

The SSRSC-diag model can be similarly solved with a standard ADMM algorithm. We performed experiments on the Hopkins155 dataset [4], and the comparison results are listed in Table 1. One can see that the variant SSRSC-diag achieves inferior performance with the original SSRSC. We have summarized the comparisons as a Table X, along with the related ablation study, in the **Page 11** (right column, colored by red) of the revised manuscript.

> ***Comment 2.4**: Any possible improvement for the impractical cube complexity?*

Method	**SSRSC** ($s = 0.5$)	**SSRSC** ($s = 0.9$)	**SSRSC-diag** ($s = 0.8$)	**RGC** [5]
Error (%)	1.53	**1.04**	1.87	1.76

Table 1: Average clustering errors (%) of SSRSC ($s = 0.5$), SSRSC ($s = 0.9$), SSRSC-diag ($s = 0.8$) and RGC on Hopkins-155 dataset [4], with the 12-dimensional data points obtained using PCA. The method RGC is provided here as suggested by **Reviewer 4**, please refer to ***Comment 4.3*** for more details.

Authors' response: Thanks for the insightful comment. The major complexity is from the cost for updating $\boldsymbol{C}$, which is $\mathcal{O}(N^3)$ when there are N data points in the data matrix $\boldsymbol{X}$. This is slow since the number N of samples in $\boldsymbol{X}$ is much larger than their feature dimension D. In order to improve the speed (while maintaining the accuracy) of SSRSC, we employ the Woodbury formula [6,7] to reduce the computational cost of updating $\boldsymbol{C}$ from $\mathcal{O}(N^3)$ to $\mathcal{O}(DN^2)$ for the inversion of the solution in Eqn. (13). Specifically, we accelerate the inversion part in the update of $\boldsymbol{C}$ in Eqn. (13) as follows:

$$(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{I})^{-1} = \frac{2}{\rho}\boldsymbol{I} - (\frac{2}{\rho})^2 \boldsymbol{X}^\top (\boldsymbol{I} + \frac{2}{\rho}\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}. \quad (4)$$

Since $(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{I})^{-1}$ is not updated during iterations, we can pre-compute it with Woodbury formula and fix it during iterations. This strategy also save plenty of computational costs. We compare the speeds of the original SSRSC described in the manuscript and the improved SSRSC (SSRSC-I) by Eqn. (4) on the tested datasets in the main paper, and the results are listed in Tables 2. One can see that the SSRSC-I achieves faster speed over the original SSRSC in most cases, and produces little influence when original SSRSC is already very fast.

Inspired by a recently published paper [8,9], we would take it as a future work to further accelerate the proposed method according to the insightful theory in [8,9]. This point is added in the "Conclusion" section (**Page 11**, right column, colored by red) as a potential future work.

Dataset	Hopkins155	YaleB			ORL	MNIST			EMNIST
		2 Subjects	5 Subjects	10 Subjects		500	2000	6000	
SSRSC	0.07	0.03	0.08	0.28	0.93	0.35	5.23	89.62	193.56
SSRSC-I	0.07	0.03	0.08	0.27	0.89	0.34	5.01	85.68	186.23

Table 2: Comparison on speed (in seconds) of SSRSC and its improved variant SSRSC-I on the five tested datasets in the main paper. For MNIST, we only show the cases of 500, 2000, or 6000 images are used.

***Comment 2.5**: Minnor comments.*

Authors' response: Thanks for the helpful suggestions.

For Q-1): We agree with the review that the title is a little bit less informative. One major confusing point may be that in standard simplex, the coefficients are summed up to 1, while in our work they are summed up to a scalar s. To more clearly reflect our idea, we have modified the title from "Simplex Representation for Subspace Clustering" to "Scaled Simplex Representation for Subspace Clustering".

For Q-2): Thanks for the suggestion. We have summarized different constraints (other than sparse or low-rank) in a new section II-C in the revised manuscript. We have described the Laplacian constraint proposed in the suggested paper there, as well as other constraints. Please refer to the “Related Work” section (**Page 3**, right column, colored by red) of the revised manuscript.

For Q-3): We have correct the broken reference. Please refer to the “References” section (**Page 11**, right column, colored by red) of the revised manuscript.

For Q-4): We are very appreciated to the suggestion, and cited the references in the “Related Work” section (**Page 3**, right column, colored by red) of the revised manuscript.

Responses to Reviewer 3

Recommendation: Prepare A Major Revision

> ***Comment 3.1**: The novelty of this paper is limited because is quite similar to the following works: “A new simplex sparse learning model to measure data similarity for clustering, IJCAI 2015”; “Robust subspace clustering via thresholding ridge regression, AAAI 2015”.*

Authors’ response: Thanks for pointing out these works. Our work has at least three differences with the two mentioned works.

First of all, the objective functions or constraints in these models are different. In our work, the proposed model with constraints is

$$\min_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda \|\boldsymbol{C}\|_F^2, \ \text{s.t.} \quad \boldsymbol{C} \geq 0, \mathbf{1}^\top \boldsymbol{C} = s\mathbf{1}^\top, \tag{5}$$

while the model of the IJCAI-2015 work is

$$\min_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2, \ \text{s.t.}\ \boldsymbol{C} \geq 0, \mathbf{1}^\top \boldsymbol{C} = \mathbf{1}^\top, \text{diag}(\boldsymbol{C}) = \mathbf{0}, \tag{6}$$

and the model of the AAAI-2015 work is

$$\min_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda \|\boldsymbol{C}\|_F^2, \ \text{s.t.}\ \text{diag}(\boldsymbol{C}) = \mathbf{0}, \tag{7}$$

One can see that, when compared with the IJCAI-2015 model (6), our model (5) is different at both the objective function and the constraints. Specifically, our model (5) employs least square regression (LSR) with the proposed scaled simplex constraints, while the IJCAI-2015 model (6). When compared with the AAAI-2015 model (7), our model (5) is very different at the constraints, the AAAI-2015 model (7) uses a diagonal constraint while our model (5) uses the proposed scaled simplex constraints, though the two models both use LSR as the objective function.

Second, the solutions to the objective functions are different. We employ standard ADMM algorithm to solve the overall model (5), the IJCAI-2015 work utilized accelerated projected gradient method to solve the model (6), while the AAAI-2015 work first projects the data into a linear space spanned by itself and then handles the errors by performing a hard thresholding operator.

Third, the objective of these works are different. In our work, we directly tackle the problems during constructing a physically reasonable affinity matrix for subspace clustering. The IJCAI-2015 work majorly deals with the issues in the Laplacian graph construction. The AAAI-2015 focus on the solving the effect of the errors (by noise) from the representation in projection space.

Since the authors of the IJCAI-2015 work called their method "simplex sparse representation", we modified the title of our manuscript as "Scaled Simplex Representation for Subspace Clustering". Since the abbreviation of "SSR" is already used in the IJCAI-2015 work, we renamed our proposed method as "scaled simplex representation based subspace clustering" (SSRSC) in the revised manuscript.

We cited the two references in the section II-C (**Page 3**, right column, colored by red) of the revised manuscript.

> ***Comment 3.2****: Despite the clustering performance, it is worthy to describe more clearly what the contributions of this paper are. "Subspace clustering via variance regularized ridge regression, CVPR 2017"; "Subspace clustering using log determinant rank approximation, KDD 2015"; "Integrating feature and graph learning with low-rank representation, Neurocomputing 2017". These methods are closely related. The first one was refereed in the paper while the last two were missed.*

Authors' response: Thanks for providing the references. They are closely related to our work, and we have cited them in the revision. The major contributions of this manuscript have been summarized in the "Introduction" section in three different aspects. To integrate the comments by the reviewers, we will describe more clearly what we did in our work.

We are very appreciated to the suggestion, and cited the references in the "Related Work" section (**Page 3**, right column, colored by red) of the revised manuscript.

> ***Comment 3.3****: Is the proposed method suitable to those data with imbalanced classes? Some discussions can be found in "Discriminative Regression Machine: A Classifier for High-Dimensional Data or Imbalanced Data".*

Authors' response: Thanks for this suggestion. The proposed method does not have assumption on the data distribution across multiple classes, and hence can be applied to those data with imbalanced classes. We note that the data with imbalanced classes in the suggested paper is used for evaluating image classification methods, not subspace clustering ones like our work. However, to validate the effectiveness of the proposed scaled simplex representation, here we modified our SSRSC model for image classification.

Denote by $\boldsymbol{y} \in \mathbb{R}^{D}$ a query data and $\boldsymbol{X} \in \mathbb{R}^{D\times N} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ the training data matrix of imbalanced classes, where $\boldsymbol{X}_k \in \mathbb{R}^{D\times N_k}$ contains the training data from class k. Given a query data $\boldsymbol{y}$ and the training data matrix $\boldsymbol{X}$, we formulate the scaled simplex representation (SSR) based classifier as follows:

$$\min_{\boldsymbol{c}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{c} \geq 0, \mathbf{1}^\top\boldsymbol{c} = s, \tag{8}$$

where $\boldsymbol{c}$ is the coding vector of $\boldsymbol{y}$ over $\boldsymbol{X}$. This model can be solved by standard ADMM algorithms as described in our manuscript, and we ignore the detailed inference process to make this response more compact. During the test process, we first apply unit ℓ_2 normalization on $\boldsymbol{y}$ and each column of $\boldsymbol{X}$, and then compute the coding vector $\hat{\boldsymbol{c}}$ according to the SSR model (8). Next we compute the class representation residual $\|\boldsymbol{y} - \boldsymbol{X}_k \hat{\boldsymbol{c}}_k\|_2$ and predict its classification, where $\hat{\boldsymbol{c}}_k$ is the coding sub-vector corresponding to the class k. We apply the modified SSR based classifier on imbalanced data classification tasks. We tune the scale s from 0 to 1 with a gap of 0.1, and SSRC achieves the best performance when $s = 0.3$. The comparisons with the DRM method are listed in Table 3, from which one can see that the SSRC achieves comparable performance with DRM.

We added the tackling of imbalanced data as a future work in the "Conclusion" section (**Page 11**, right column, colored by red) of the revised manuscript. The suggested paper is cited as an important reference.

Dataset	haberman	ecoli1	glass4	new thyroid 2	ecoli2	glass6	yeast3	ecoli3	glass2	ecoli4
DRM(P)	0.6129	0.8728	0.9219	0.9852	0.9057	0.9323	0.8860	0.8668	0.7370	0.9343
SSRC	0.6013	0.8691	0.9187	0.9734	0.9013	0.9278	0.8795	0.8712	0.7120	0.9189

Table 3: Comparison on average G-mean results of the DRM and our modified SSRC on classification of 10 imbalanced datasets used in [10].

> ***Comment 3.4***: *Any theoretical guarantee on the convergence of the proposed algorithm?*

Authors' response: Since the proposed SSRSC model is convex, and it only has two variables in the ADMM algorithm, the proposed algorithm can be guaranteed to converge to a global optimal solution. To illustrate this point, we plot the convergence curve in Figure 2. It can be seen that the proposed algorithm converges quickly in several iterations. We added the figure of convergence curves and the corresponding convergence analysis, at the end of Section III-B (**Page 5**, left column, colored by red).

> ***Comment 3.5***: *It appears straightforward to extend the proposed method to nonlinear version by using kernel approach. Why did not the authors do this?*

Authors' response: Thanks for the insightful suggestion. The major goal of our work is to develop an effective scaled simplex representation (SSR) for subspace clustering. Extensive experiments on several tasks have demonstrated the effectiveness of SSR. We will consider to extending it to nonlinear version as a future work. We are very appreciated to add this point as a potential future work in the "Conclusion" section (**Page 11**, right column, colored by red).

Responses to Reviewer 4

Recommendation: Prepare A Major Revision

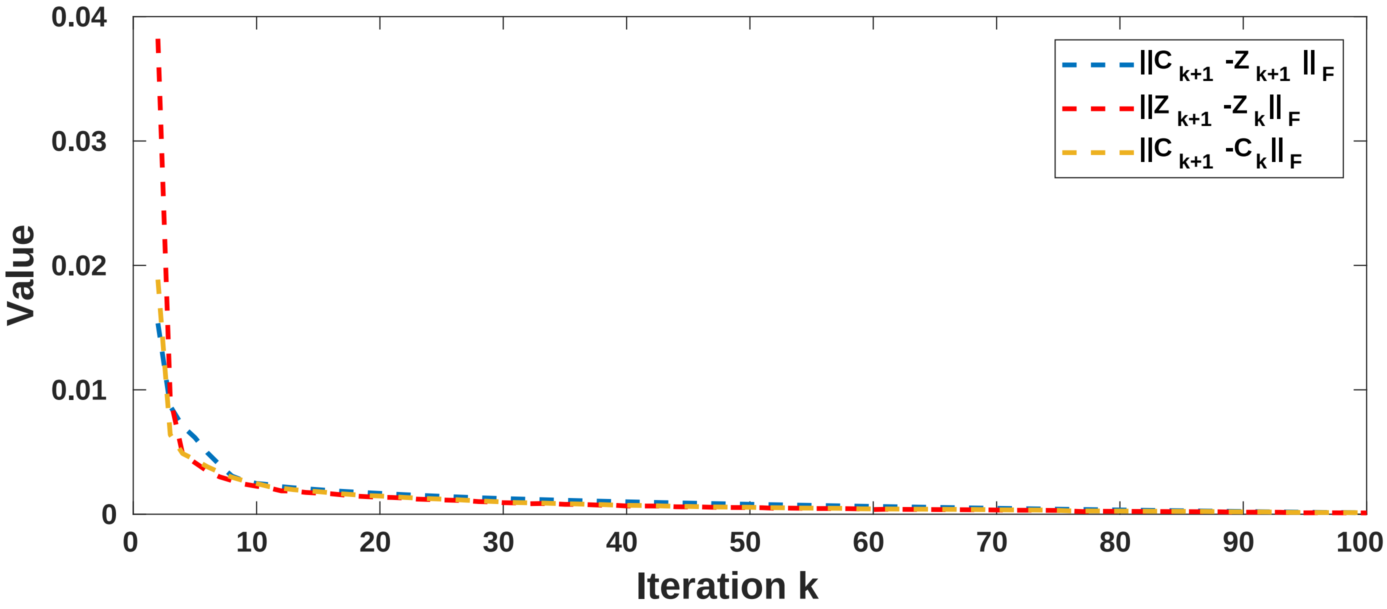


Figure 2: The convergence curves of $\|\mathbf{C}_{k+1}-\mathbf{Z}_{k+1}\|_F$ (red line), $\|\mathbf{Z}_{k+1}-\mathbf{Z}_k\|_F$ (blue line), and $\|\mathbf{C}_{k+1}-\mathbf{C}_k\|_F$ (green line) of the proposed SSRSC on the "1R2RC" sequence from the Hopkins155 dataset [4].

> ***Comment 4.0****: Some recent work in the literature has adopted the nonnegative constraint, e.g., "Kernel-driven Similarity Learning, Neurocomputing, 2017"; "Low-rank Kernel Learning for Graph-based Clustering, Knowledge-based Systems, 2019".*

Authors' response: Although this is not a question, we still list it here to thank the reviewer for providing us two closely related work. Inspired by these paper, we also found another work [11] closely related to ours. We have cited and commented these work in a new section II-C of the *Related Work* section (**Page 3**, right column, colored by red) in the revised manuscript. Also, to make the *Related Work* section as compact as possible, we remove the description of iterative methods, algebraic methods, and statistical methods in Section II-A (**Page 2**, right column, colored by red) of the revised manuscript, since they are not very closed related to our work, and already described in previous paper.

> ***Comment 4.1****: the authors use different comparison methods for different data sets. For example, why is DSC and ESC only used in Extended Yale B dataset and MNIST, respectively?*

Authors' response: Thanks for the question. We directly copied from the corresponding paper the results of DSC and ESC on the ORL and EMNIST datasets, respectively. The main reasons we do not report their results on other datasets are twofold: 1) DSC need to be trained on specific datasets, but it has some bugs when being trained or tested; 2) the clustering errors of ESC are unreasonably high on the datasets tested in our manuscript. To provide more comparisons while not confusing the reader, we only provide their results on the datasets used in our manuscript, and do not report the results of these methods on other datasets. But on all tested datasets, we compared with nine competing subspace clustering methods.

> ***Comment 4.2***: *the authors say that they use s=0.5 in all experiments in subsection A (Implementation Details). However, they claim different values for each data set later. What's the reason?*

Authors' response: In SSRSC, the scalar s is fixed as $s = 0.5$ to balance its performance on all datasets in our experiments. However, the SSRSC with different scalar s can achieve better performance on specific dataset. For example, SSRSC with $s = 0.9$ achieves 1.04% on the Hopkins155 dataset. On the other datasets, SSRSC with different s can also achieve lower accuracies than SSRSC with $s = 0.5$.

> ***Comment 4.3***: *in subsection A (Proposed SSRSC Model), the authors claim that it is beneficial to get rid of the diagonal constraint, especially when the data is noisy. However, no experiments validate this. For this, it would be interesting to compare with another type of clustering method, "Robust Graph Learning from Noisy Data, IEEE Cybernetics, 2019".*

Authors' response: Thanks for the insightful suggestion. We agree with the reviewer that we should validate the argument. To achieve this, we compare with the suggested method on the Hopkins155 dataset, together with the SSRSC model with additional diagonal constraint of diag$(\boldsymbol{C}) = \mathbf{0}$ (we call this method *SSRSC-diag*). The results are listed in Table 1 (please refer to ***Comment 2.3*** of **Reviewer 2** for more details). One can see that the variant *SSRSC-diag* achieves inferior performance with the original SSRSC. The suggested RGC achieves slightly better performance than *SSRSC-diag*, but is still inferior to our SSRSC without diagonal constraint. We have summarized the comparisons as a Table X, along with the related ablation study, in the **Page 11** (right column, colored by red) of the revised manuscript.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[3] C. Lu, H. Min, Z. Zhao, L. Zhu, D. Huang, and S. Yan. Robust and efficient subspace segmentation via least squares regression. In *ECCV*, pages 347–360. Springer, 2012.

[4] R. Tron and R. Vidal. A benchmark for the comparison of 3d motion segmentation algorithms. In *CVPR*, 2007.

[5] Z. Kang, H. Pan, S. C. H. Hoi, and Z. Xu. Robust graph learning from noisy data. *IEEE Transactions on Cybernetics*, pages 1–11, 2019.

[6] K. Riedel. A sherman-morrison-woodbury identity for rank augmenting matrices with application to centering. *SIAM Journal on Matrix Analysis and Applications*, 13(2):659–662, 1992.

[7] N. J. Higham. *Accuracy and stability of numerical algorithms*, volume 80. Siam, 2002.

[8] C. You, C. Li, D. P. Robinson, and R. Vidal. Scalable exemplar-based subspace clustering on class-imbalanced data. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[9] C. Peng, C. Chen, Z. Kang, J. Li, and Q. Cheng. Res-pca: A scalable approach to recovering low-rank matrices. In *CVPR*, June 2019.

[10] C. Peng and Q. Cheng. Discriminative regression machine: A classifier for high-dimensional data or imbalanced data, 2019.

[11] Z. Kang, H. Xu, B. Wang, H. Zhu, and Z. Xu. Clustering with similarity preserving. *Neurocomputing*, 2019.