REPLICATING THE INTERPRETABLE KERNEL DIMENSION REDUCTION PROBLEM USING ITERATIVE SPECTRAL METHOD

Han Zhou
McGill University
han.zhou@mail.mcgill.ca

Hao Shu
McGill University
hao.shu@mail.mcgill.ca

Gengyi Sun
McGill University
gengyi.sun@mail.mcgill.ca

December 15, 2019

ABSTRACT

Data processing has always been one of the most popular approaches that data scientists used to improve the performance of the models. Among all, the kernel method for support vector machine is a powerful technique that can significantly reduce the dimensional complexity of the input features. The standard approach that maps the dataset to a high dimensional space before the projection of the dataset is strong at capturing the non-linear relationship among features but made the feature not interpretive anymore. To solve the problem, an Interpretive Kernel Dimension Reduction method is proposed. However, this method requires a non-convex manifold that is hard to solve. In the paper recently published in NeurIPS 2019, C.Wu and his team have claimed a break though which extends the theoretical guarantee of the Iterative Spectral Method(ISM), which was originally used solely for Gaussian kernel in alternative clustering, to the entire family of kernels. Besides, they have proved that this wide-ranged IKDR method can also be applied to all learning paradigms with an outstanding performance compared with other commonly used IKDR methods such as Dimension Growth and Steifel Manifold. To reproduce their work, we proposed an experimental approach to their claimed baseline and result. We first concluded the three most important examine criteria from their claimed contributions. Then we proceed with the experiment by examining each criterion we concluded and checked if they met the claimed baseline. After a series of cautious experiments and reproducing. We have confirmed the correctness of their work, hence validate the newly established baseline. Not only verifying their work, but we have also discovered the trade-off between reducing dimension and improving accuracy, hence built up a more insightful knowledge towards this area of research.

1 Introduction

The researching area of the paper to be reproduced is mainly focusing on the optimization of kernel construction used in machine learning classifiers like support vector machine. Kernel is a powerful technique for its use in significantly reducing the dimension of input datasets, leading to a reduction in the complexity of features and run-time [1] [2]. This reduction was often gained from mapping the dataset to a high dimensional feature space, then projecting it onto a low dimensional space [3]. This approach can easily capture the non-linear relationship with the most dominant eigenvectors of the kernel matrix [4], hence, the established solution. However, this process also makes the transformation not interpretable. The opposite approach which projects the dataset onto a low dimensional space first then maps it to a high dimensional space provides information on how the original features linearly related to the transformed features [5]. From this insight, C. Wu et al.[6] refer all the formulation which increases the Hilbert Schmidt Independence Criterion(HSIC)[7] as Interpretable Kernel Dimension Reduction(IKDR). Wu et al.[6] found that the Iterative Spectral Method used for alternative clustering greatly reduced the run-time for a problem that took 2 days for Dimension Growth. Later the team expands ISM's [8] theoretical guarantees to all kernels and more learning paradigms. In terms of replicating their experiment, Shu, Sun, and Zhou validate the liability of their algorithm and examine the quantity of reduction in feature dimension, run-time, and HSIC. We also evaluate the strength of this algorithm on different learning paradigms(supervised and unsupervised).

2 Related Work

Numerous methods have been proposed for dimension reduction of input features in machine learning. It belongs to a category of optimization on a manifold due to its orthogonality constraint [1], which can be modeled as a Grassmann manifold [9] or Stiefel manifold [10] [11][12]. In 2009, Theis, Cason, and Absil [13] employed a trust-region method for minimizing the cost function on the Stiefel manifold and obtained a higher numerical efficiency. In 2010, Wen and Yin [14] proposed to unfold the Stiefel manifold into a flat plane to preserve the constraints. In 2011, Turaga et al. [9] demonstrated the improved performance of Grassmann manifold in a wide variety of vision applications such as activity-based video clustering. Later, Boumal and Absil [15] recast the dimension reduction problem as an unconstrained optimization problem on the Grassmann manifold, and apply first- and second-order Riemannian trust-region methods to solve it. Besides the manifold methods, Niu, Dy, and Jordan [16] proposed Dimension Growth (DG) in 2011, which automatically learns the relevant dimensions and spectral clustering simultaneously. Dimension Growth is demonstrated to yield significant improvements in the performance of spectral clustering.

In 2018, Wu et al. [8] proposed the Iterative Spectral Method (ISM) for alternative clustering, their experiments demonstrated that ISM has a much higher execution speed with much lower objective cost compared to Dimension Growth. Wu et al. also presented ISM as an approach to Kernel Dimension Reduction problems across several learning paradigms, including supervised dimension reduction [17][18], semi-supervised dimension reduction [19], and unsupervised dimension reduction [20]. The experimental evidence [8] proves the excellent performance of ISM on a wide range of learning paradigms in the kernel and demonstrate its efficiency in terms of the low time complexity and lower objective cost compared to competing approaches.

However, ISM can only theoretically guarantee for the Gaussian kernel [8]. Since any kernel method using ISM is now limited to the Gaussian kernel, the ISM's usage is greatly limited. Then in 2019, Wu et al. [6] extended the theoretical guarantees of ISM to an entire family of kernels, thereby empowering ISM to solve any kernel method of the same objective.

3 Dataset and Setup

	Abalone	Iris	Reddit	Wine	Cancer
Data Size	4057	130	1800	1449	684
Number of Features	8	4	10438	11	9
Output categories	15	3	2	2	2

Table 1: Datasets Used

Multiple datasets were used to examine the correctness of the work. Two of them were selected by us from the UCI website [21], we picked 2 most popular datasets, one named Iris with few features and another named abalone with a normal number of features for comparison regarding the performance of dimensional reduction of ISM in different scenarios.

Another dataset was acquired from this semester's study, a Reddit comment collection [22] which consists of 1800 real Reddit comments from 2 different categories indicating which subreddit each comment belongs to. Nevertheless, we chose the two publicly opened datasets, the wine dataset, and the breast cancer dataset, mentioned in their paper[6] to check the consistency of their work.

3.1 Pre-processing

For the abalone dataset, the iris dataset, the wine quality dataset and the cancer dataset, we directly use the raw data to check their performance of the optimization.

The Reddit comments came from a previous assignment was used. As the assignment used to have 20 different categories of comments, we decide to randomly pick two subreddits, in this case, Canada and Europe. Extracted a small subset of each class, then shuffled the order. Since the raw data are plain text as Reddit comments, we used count vectorizer counting for the occurrence for words, then we used TF*IDF transformer transforming them into a matrix of dimension with the number of unique words.

4 Proposed Approach

4.1 Examine Criteria

In this paper, authors have claimed that they managed to achieve the following contributions based on current progress:

- Generalize the theoretical guarantees of ISM from Gaussian kernel to the entire family of kernel by found a general formula for generating matrix pairs of any kernel within the family.
- Demonstrate the efficiency of ISM comparing to other optimization techniques in various IKDR learning paradigms. In this paper, they are compared with Dimension Growth, the Stiefel Manifold approach, and the Grassmann Manifold in terms of computational cost and time cost.
- Highlight the capability of ISM on achieving better run-time and objective performance in the usage of matrix or linear combination of matrices in place of kernels.

In our reproduction, we decided to valid and reproduce their progress by examining these criteria:

- Validate the extending and generalizing of the ISM in different kernels and various learning paradigms.
- Demonstrate the functional capability of this optimization in terms of maintaining the non-linear dependence between feature and output while reducing the dimension of input features.
- Verify the reduction of computation cost and time comparing to other optimization methods.

4.2 Reproduction Design

The reproduction is designed according to the criteria specified above. An experimental approach is used to verify the correctness of their works, The preset datasets were run on different kernel sets, with different learning paradigms, and have their HSIC rating compared with the original datasets' rating, as well as the results generated by other optimization models. The kernels we will be using in the experiment are: Linear, Polynomial, Squared, and Gaussian. The dataset used, as introduced above, are: iris, abalone, wine, cancer, and simplified Reddit data. The other optimization methods compared are: Steifel Manifold (SM) and Grassmann Manifold (GM) from **Pymanopt** package [23]. After the dataset is processed, a systematical analysis will be performed concerning each criterion by selecting proper combinations of results.

5 Results

1. Validation of implementing different kernels in terms of dimensional reduction while maintaining proper non-linear dependence.

Here the HSIC is used as an indicating parameter for the increase in the non-linear correlation between the input feature and output result.

Kernel Type	Gaussian	Linear	Polynomial	Squared
Input Dimension	4057 * 8	4057 * 8	4057 * 8	4057 * 8
Output Dimension	4057 * 5	4057 * 5	4057 * 5	4057 * 5
Train Time	110.412	0.547	1.433	1.802
Initial HSIC	147857.5967	3511111.717	1129198.45	7022223.434
Final HSIC	160253.5296	3510969.164	2047491.924	7021938.328
Train Accuracy	0.2822	0.2938	0.2926	0.2938
Test Accuracy	0.2	0.1833	0.175	0.1833

Table 2: Abalone Dataset Using Different Types of Kernel

Kernel Type	Gaussian	Linear	Polynomial	Squared
Input Dimension	1800*10438	1800*10438	1800*10438	1800*10438
Output Dimension	1800*1449	1800*1449	1800*1449	1800*1449
Train Time	468.000	355.053	448.683	678.468
Initial HSIC	457.7906	10594180.0347	3678.4982	21188360.0694
Final HSIC	23305.0655	10594180.0347	99503.2065	21188360.0694
Train Accuracy	1.0000	0.9989	0.9994	0.9967
Test Accuracy	0.5200	0.6950	0.6500	0.6850

Table 3: Reddit Comments Using Different Types of Kernel

As shown in the tables, for each kernel, ISM achieved a significant reduction in feature dimension while kept a minimum effect in HSIC, in case of Gaussian and polynomial, there's a significant improvement with respect to HSIC, which we supposed was due to the nature of the dataset we chose. Hence, the theoretical expansion of ISM from Gaussian kernel to other kernels in the family has been verified.

2. Parallel comparison of performances between different optimization methods.

	Time Used(sec)	accuracy(%)
SM	233	0.216
GM	198	0.217
ISM	1.2	0.220

Table 4: Comparing ISM with two other optimizer in terms of time cost

We compared the run time of ISM with Steifel Manifold and Grassmann Manifold method using the Iris dataset, a relatively small dataset (containing only 130 data points). As shown in the table above, ISM has an outstanding time efficiency compared with two other methods used in the paper.

3. Checked the consistency of performance with respect to different IKDR learning paradigms.

Kernel Type	Gaussian	Linear	Polynomial	Squared
Input Dimension	1449 * 11	1449 * 11	1449 * 11	1449 * 11
Output Dimension	1449 * 9	1449 * 9	1449 * 9	1449 * 9
Train Time	1.118	0.568	1.074	1.981
NMI	0.86	0.85	0.84	0.85

Table 5: Unsupervised Learning on Wine Dataset

To investigate the performance of ISM in different IKDR learning paradigms, we used spectral clustering as an example of unsupervised learning, which is the same as the paper used. Two datasets that they have used in the paper were run to check the result consistency, and one of them is shown in the table above. The examine criteria we used are normalized mutual information(NMI), which is the same criteria in their report.

The result shows that their method pretty much fulfilled their claims in terms of dimension reduction and time efficiency, as most of the result we get from the wine and cancer data set shows similar NMI rate as they claimed. The time cost is, however, a little different than their report, which we suppose the bias may due to the difference in experiment devices.

	Accuracy With Kernel[1]	Accuracy Without Kernel[2](%)
Reddit	0.6375	0.7300
Abalone	0.1854	0.2417

Table 6: Comparing Test Accuracy With Kernel and Without Kernel

[1] This is calculated by averaging the test scores of the 4 types of kernel.
[2] This accuracy refers to the cross-validation accuracy using support vector machine.

One thing worth noticing is that the reduction in input feature dimension costs expensively in test accuracy. Particularly in Reddit comments classification: since all kernel tricks significantly reduced the dimensions of the features of the Reddit comments. As a remark, the inputs of Reddit comments are transformed from raw text to a matrix of integers representing the TF*IDF property. All four kernels reduce the number of input features from 10438 to 1449, thus the information is lost during the feature reduction process. The loss information hurts the final accuracy because a huge amount of keywords determining the correct subreddit are lost. We supposed that this loss was due to the low mutual information nature of the dataset used. This indicates the limitation and constraints of this method during application.

6 Discussion and Conclusion

Overall, this paper has successfully demonstrated their contribution to the Iterative Spectral Method, it has extended the theoretical guarantee from solely Gaussian kernel to the entire kernel family and present the reliable experimental evidence to support their result. During the reproduction, we have successfully reproduced most of their claimed achievements, and discovered some properties regarding to the ISM in terms of optimization. The ISM is proved to be a strong and efficient tool in solving the IKDR problems, however, it has its limitation and constraints. This reproduction help us understanding the ways of implementing general optimization method and the limitation to this process. The new baseline has been established as the test data claimed in their paper, and verified by our reproduction.

7 Future Work

The IKDR does make a good job at reducing the input dimensional complexity, but the reduction process also hurts the prediction accuracy. Further development can be directed into finding a way that reducing the kernel dimension with less trade off in the prediction accuracy.

8 Statement of Contribution

All members have made significant contributions towards this project. The amount of work for each member is described as follows:

Han Zhou : Report Write-up, Testing Abalone, Iris and Wine Dataset, Code Writing

Gengyi Sun : Report Write-up, Testing Reddit Comments Dataset, Code Writing.

Hao Shu :Report Write-up, Report Formatting.

References

[1] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol. 10, pp. 1299–1319, July 1998.

[2] B. Schölkopf, A. Smola, and K. Muller, "Kernel principal component analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1327 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 583–588, Springer Verlag, 1997.

[3] D. Niu, J. Dy, and M. I. Jordan, "Dimensionality reduction for spectral clustering," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (G. Gordon, D. Dunson, and M. Dudík, eds.), vol. 15 of *Proceedings of Machine Learning Research*, (Fort Lauderdale, FL, USA), pp. 552–560, PMLR, 11–13 Apr 2011.

[4] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[5] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *Ann. Statist.*, vol. 37, pp. 1871–1905, 08 2009.

[6] C.-T. Wu, J. Miller, Y. Chang, M. Sznaier, and J. Dy, "Solving interpretable kernel dimension reduction," *ArXiv*, vol. abs/1909.03093, 2019.

[7] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Algorithmic Learning Theory* (S. Jain, H. U. Simon, and E. Tomita, eds.), (Berlin, Heidelberg), pp. 63–77, Springer Berlin Heidelberg, 2005.

[8] C. Wu, S. Ioannidis, M. Sznaier, X. Li, D. Kaeli, and J. Dy, "Iterative spectral method for alternative clustering," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (A. Storkey and F. Perez-Cruz, eds.), vol. 84 of *Proceedings of Machine Learning Research*, (Playa Blanca, Lanzarote, Canary Islands), pp. 115–123, PMLR, 09–11 Apr 2018.

[9] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on grassmann and stiefel manifolds for image and video-based recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2273–2286, Nov 2011.

[10] A. Edelman, T. Arias, and S. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, 06 1998.

[11] I. M. James, *The topology of Stiefel manifolds*, vol. 24. Cambridge University Press, 1972.

[12] Y. Nishimori and S. Akaho, "Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold," *Neurocomput.*, vol. 67, pp. 106–135, Aug. 2005.

[13] F. J. Theis, T. P. Cason, and P. A. Absil, "Soft dimension reduction for ica by joint diagonalization on the stiefel manifold," in *Independent Component Analysis and Signal Separation* (T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, eds.), (Berlin, Heidelberg), pp. 354–361, Springer Berlin Heidelberg, 2009.

[14] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, pp. 397–434, Dec. 2013.

[15] N. Boumal and P. antoine Absil, "Rtrmc: A riemannian trust-region method for low-rank matrix completion," in *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds.), pp. 406–414, Curran Associates, Inc., 2011.

[16] D. Niu, J. G. Dy, and M. I. Jordan, "Iterative discovery of multiple alternativeclustering views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1340–1353, July 2014.

[17] M. Masaeli, G. Fung, and J. G. Dy, "From transformation-based dimensionality reduction to feature selection," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, (USA), pp. 751–758, Omnipress, 2010.

[18] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357 – 1371, 2011.

[19] M. J. Gangeh, S. M. A. Bedawi, A. Ghodsi, and F. Karray, "Semi-supervised dictionary learning based on hilbert-schmidt independence criterion," in *Image Analysis and Recognition* (A. Campilho and F. Karray, eds.), (Cham), pp. 12–19, Springer International Publishing, 2016.

[20] M. Wang, F. Sha, and M. I. Jordan, "Unsupervised kernel dimension reduction," in *Advances in Neural Information Processing Systems 23* (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds.), pp. 2379–2387, Curran Associates, Inc., 2010.

[21] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[22] "Reddit comment classification comp 551." https://www.kaggle.com/c/reddit-comment-classification-comp-551/data. Accessed: 15- Dec- 2019.

[23] J. Townsend, N. Koep, and S. Weichwald, "Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation," *Journal of Machine Learning Research*, vol. 17, no. 137, pp. 1–5, 2016.