Feature Reduction Based on Sum-of-SNR (SOSNR) Optimization

FEATURE REDUCTION BASED ON SUM-OF-SNR (SOSNR) OPTIMIZATION

Yinan Yu[a,b], *Tomas McKelvey*[a], *S.Y. Kung*[b]

Chalmers University of Technology [a]
Gothenburg, Sweden
{yinan,tomas.mckelvey}@chalmers.se

Princeton University [b]
Princeton, NJ, USA
{yinany, kung}@princeton.edu

ABSTRACT

Dimensionality reduction plays an important role in machine learning techniques. In classification, data transformation aims to reduce the number of feature dimensions, whereas attempts to enhance the class separability. To this end, we propose a new classifier-independent criterion called "Sum-of-Signal-to-Noise-Ratio" (SoSNR). A framework designed for maximization with respect to this criterion is presented and three types of algorithms, respectively based on (1) gradient, (2) deflation and (3) sparsity, are proposed. The techniques are conducted on standard UCI databases and compared to other related methods. Results show trade-offs between computational complexity and classification accuracy among different approaches.

Index Terms— Sum-of-SNR, feature reduction, classification, Fisher's Score, SODA

1. INTRODUCTION

Given a fixed number of training samples, the computational complexity and generalization performance of machine learning algorithms commonly depend on the number of input variables, i.e. the features.

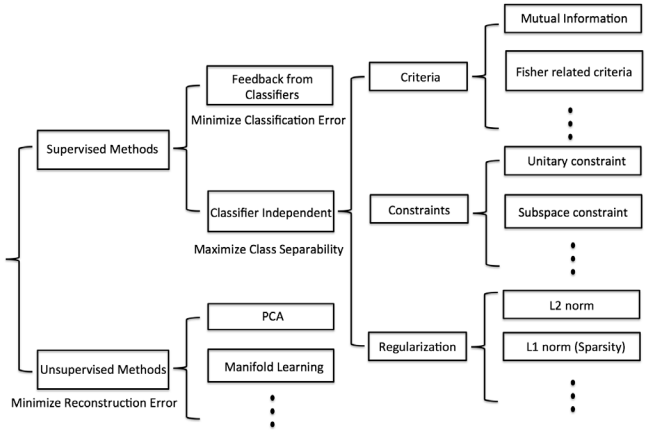

Fig. 1. A brief overview of related feature reduction methods.

The problem of feature reduction [3, 4, 6] is stated as follows. Given a p dimensional input variable $\mathbf{x} = (x_1, \cdots, x_p)$, we would like to find a k dimensional ($k < p$) representation $\mathbf{x}' = (x'_1, \cdots, x'_k)$ with respect to *some criteria*, such that the valid information contained in the original feature space is captured. Figure 1 provides a brief overview of existing approaches, which essentially fall into two categories with respect to their objectives:

I Unsupervised methods[2]: evaluating the performance by the reconstruction error.

II Supervised methods: (a) minimizing the classification error rate, or (b) maximizing the class separability, where the techniques depend on either the feedback from the equipped classifier (classifier-dependent), or maximizing a cost function without specifying the classification criterion (classifier-independent). Among all the objective functions, there are two frequently used criteria: the mutual information and Fisher score [8, 10]. Penalties are commonly applied for the sake of robustness [22, 11] or the sparsity [19, 17, 20] of the feature vectors.

In this paper, we mainly focus on 'classifier-independent' approaches introduced in catalog II (b). In general, this type of techniques provide a more flexible and adaptive solution. A new criterion as a measure of class separability is proposed and discussed from the viewpoint of Signal-to-Noise Ratio (SNR) with respect to Fisher's discriminant criteria. The objective function is formulated as a Sum-of-SNR (SoSNR) optimization problem subject to orthogonality constraints.
This paper emphasizes on three aspects:

(a) SoSNR as a new classifier-independent criterion for feature reduction is proposed and the corresponding numerical analysis is presented.

(b) The relation between SoSNR and classification error is exploited in order to verify the validity of the new criterion, i.e. SoSNR increases with respect to a better classification performance for various techniques.

(c) The classification performance is evaluated for different reduced feature dimension k's. For a given error rate, techniques with lower 'compression rate' $\frac{k}{p}$ are preferable.

2. CLASSIFIER-INDEPENDENT CRITERION: SUM-OF-SNR (SOSNR)

In this section, we propose a classifier-independent criterion for measuring the 'overall distance' between classes.

Definition (SoSNR)**.** *Given data*

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\, y_i \in \{-1, 1\}\} \tag{1}$$

and a map φ*:* $\mathbb{R}^p \to \mathbb{R}^k$ *with* $k < p$*, Sum-of-SNR (SoSNR) is defined as*

$$SoSNR = \sum_{i=1}^{k} \frac{s_i^2(\varphi; \mathcal{D})}{n_i^2(\varphi; \mathcal{D})} \tag{2}$$

where $s(.)$ *and* $n(.)$ *are predefined functions for computing the between class distance (signal) and within class distance (noise) for the given data set;* $s(.)_i$ *and* $n(.)_i$ *denote the* i^{th} *dimension of the corresponding vectors.* □

Using this criterion, the optimization problem is formulated as follows. Given training data $\mathcal{D}$ and $k < p$, find a functional φ: $\mathbb{R}^p \to \mathbb{R}^k$ such that

$$\varphi^* = \arg\max_{\varphi} \quad \left\{ \sum_{i=1}^{k} \frac{s_i^2(\varphi;\mathcal{D})}{n_i^2(\varphi;\mathcal{D})} \middle| \varphi \in S \right\} \tag{3}$$

where S represents the searching space.

This framework defines a sum-of-ratio [1] optimization problem, where the solutions can be found or approximated using fractional programming techniques. Sum-of-ratio optimization is widely used in many domains, where the numerators and denominators may represent various forms of "profits" and "costs". Though they are known to be complex to solve, the problems could be simplified by adopting special forms.

The framework offers three favorable attributes:

- Effectiveness: in practice, SoSNR is highly correlated to the classification performance, which is empirically verified in Section 4.
- Simplicity: as a data driven approach, no probability distribution model is mandatory.
- Flexibility: the definition of signal $s^2(.)$, noise $n^2(.)$ and the searching space S can be adapted to different scenarios.

Special case study: FDA-type SNR

The most popular SNR is perhaps the well known Fisher's Score, where the ratio between $s(.)^2$ and $n(.)^2$ is used to measure the class separability. This leads to the Fisher Discriminant Analysis (FDA) [16].

Given class label $c \in \{+,-\}$ and training data $\mathcal{X}_c = \{\mathbf{x}_1^c, \cdots, \mathbf{x}_{N_c}^c\}$, the class separation is measured using the 'between-class scatter matrix' $\mathbf{S}_B$ and the 'within-class scatter matrix' $\mathbf{S}_W$, which are respectively defined as:

$$\begin{aligned} \mathbf{S}_W &= \sum_{c\in\{+,-\}} \sum_{i=1}^{N_c} (\mathbf{x}_i^c - \mu^c)(\mathbf{x}_i^c - \mu^c)^T \\ \mathbf{S}_B &= (\mu^+ - \mu^-)(\mu^+ - \mu^-)^T \end{aligned}$$

where μ^+ and μ^- are the mean vector estimated from the corresponding class $+$ and $-$; and throughout the paper, we assume that $\mathbf{S}_W$ has full rank. Fisher's Score is defined as

$$J = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{4}$$

Fisher's Score provides a classical measure of the between-class distance (signal) and within-class distance (noise). One illustrative example can be visualized in Figure 2, where the comparison shows how the class separability depends on different 'signal' and 'noise' level.

By applying Equation (4) to (2) we have:

$$SoSNR = \sum_{i=1}^{k} \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i}{\mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i} \tag{5}$$

In this paper, we focus on the development based on Equation (10). For simplicity, SoSNR is defined based on Equation (5) instead of its generic definition in Equation (2) throughout this paper.

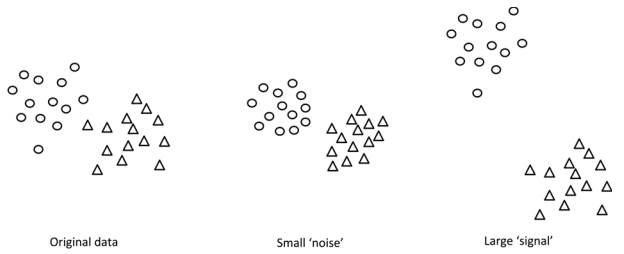

Fig. 2. Demonstration of small 'noise' and large 'signal'. This can be interpreted using the concept of between-class scatters and within-class scatters.

3. NUMERICAL ALGORITHMS

3.1. Gradient Based Algorithm

3.1.1. Linear learning model

In FDA, the Fisher's Score is maximized to obtain a vector $\mathbf{w}$ such that Equation (4) is maximized. The solution vector is the generalized eigenvector corresponding to the largest eigenvalue of the problem $\mathbf{S}_B\mathbf{w} = \mathbf{S}_W\mathbf{w}\lambda$. This concept has also been extended to feature reduction techniques [18], which means a $p \times k$ matrix is produced instead of a $p \times 1$ vector. However, given number of classes C, existing techniques are essentially solving an eigenvalue problem and therefore only $C-1$ vectors can be found due to the rank deficiency of matrix $\mathbf{S}_B$. Instead, in SoSNR optimization, we have the following.

Formulation (SoSNR Optimization)**.**

$$\begin{aligned} &\underset{\mathbf{w}_1,\cdots,\mathbf{w}_k}{\textit{maximize}} \quad \sum_{i=1}^{k} \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i}{\mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i} \\ &\textit{subject to:} \quad \mathbf{w}_i^T \mathbf{w}_i = 1, \ \ \mathbf{w}_i^T \mathbf{w}_j = 0 \ \ (\forall i \neq j). \end{aligned} \tag{6}$$

From the primal form, we can compute the Lagrangian as follows

$$\begin{aligned} \mathcal{L} &= \min - \{ \sum_{i=1}^{k} \left(\frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i}{\mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i} \right) \\ &+ \sum_{j=1}^{k} \alpha_j \left(\mathbf{w}_j^T \mathbf{w}_j - 1 \right) + \sum_{m=1}^{k} \sum_{l>m}^{k} \beta_l \gamma_m \left(\mathbf{w}_l^T \mathbf{w}_m \right) \} \end{aligned} \tag{7}$$

which results in $\frac{k(k+1)}{2} + k(p+2)$ variables growing quadratically with respect to k. Optimality is achieved when

$$\begin{aligned} \nabla_{\mathbf{w}_i} \mathcal{L} &= (2\mathbf{S}_B \mathbf{w}_i) \mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i - 2\mathbf{S}_W \mathbf{w}_i \mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i \\ &\quad + 2\alpha_i \mathbf{w}_i \left(\mathbf{w}_i^T \mathbf{S}_w \mathbf{w}_i \right)^2 + \beta_j \gamma_i \sum_{j>i}^{k} \mathbf{w}_j \left(\mathbf{w}_i^T \mathbf{S}_w \mathbf{w}_i \right)^2 = 0 \\ \nabla_{\alpha_i} \mathcal{L} &= \mathbf{w}_i^T \mathbf{w}_i - 1 = 0, \ \ \nabla_{\beta_i} f = \sum_{i=1}^{k} \gamma_i \mathbf{w}_i^T \mathbf{w}_j = 0 \\ \nabla_{\gamma_i} \mathcal{L} &= \sum_{j>i}^{k} \beta_j \mathbf{w}_i^T \mathbf{w}_j = 0, \quad \forall i, j \neq i. \end{aligned}$$

Maximization of SoSNR with respect to orthogonal constraints is a nonconvex and nonlinear optimization problem. With a properly chosen initial point, nonlinear programming techniques can be applied to obtain local optimal solutions. Monte Carlo simulations offer us an opportunity of searching the best result among local optima. In this work, a classical gradient based optimization technique called interior-point method is employed. It is one of the most popular approaches for solving nonlinear optimization problems[9].

3.1.2. Kernel based learning model

Kernel method [14, 15] is one of the most powerful techniques of modeling nonlinear functions. Define $\phi(\mathbf{x}_i)$'s the vectors in the intrinsic space and $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_N)]$. Let matrices $\mathbf{M}$ and $\mathbf{N}$ [13]:

$$\mathbf{M} = (\mathbf{M}_+ - \mathbf{M}_-)(\mathbf{M}_+ - \mathbf{M}_-)^T. \tag{8}$$

$$\mathbf{N} = \sum_{c\in\{+,-\}} \mathbf{K}_c(\mathbf{I} - \frac{1}{N_c}\mathbf{E})\mathbf{K}_c^T, \tag{9}$$

where the row vectors of $\mathbf{M}_c$ are written as $(\mathbf{M}_c)_j = \frac{1}{N_c}\sum_{t=1}^{N_c} k(\mathbf{x}_j, \mathbf{x}_t^c)$. Matrix $\mathbf{K}_c$ is the kernel matrix $\mathbf{K}_c = \mathbf{\Phi}^T\mathbf{\Phi}_c$ and $\mathbf{E}$ is a matrix with all ones as its entries. Furthermore, denote $k(\mathbf{x}) = \mathbf{\Phi}^T\phi(\mathbf{x})$, we have

Formulation (Kernel SoSNR Optimization)**.** *Given training data $\mathcal{D}$ and $k < p$, find optimal vectors $\mathbf{a}_1 \cdots \mathbf{a}_k$, such that:*

$$\begin{aligned} \underset{\mathbf{a}_i}{\textit{maximize}} \quad & \sum_{i=1}^{k} \frac{\mathbf{a}_i^T\mathbf{M}\mathbf{a}_i}{\mathbf{a_i}^T\mathbf{N}\mathbf{a}_i} \\ \textit{subject to:} \quad & \mathbf{a}_i^T\mathbf{a}_i = 1,\ \ \mathbf{a}_i^T\mathbf{K}\mathbf{a}_{j\neq i} = 0, \mathbf{a}_i^T\mathbf{N}\mathbf{a} > 0 \end{aligned} \tag{10}$$

Since matrix $\mathbf{N}$ is in general rank deficient, we need the inequality constraint to prevent degeneration. Similar to Equation (6), there are $\frac{k(k+1)}{2} + k(p+3)$ variables involved in the optimization problem.

3.2. Deflation Based Algorithm

3.2.1. SODA

The previous algorithms intend to tackle Equation (6) directly, where large amount of computations are involved. Alternatively, we revisit a recently proposed dimensionality reduction technique called Successively Orthogonal Discriminant Analysis (SODA) [5], which could be considered as an approximation to the SoSNR Optimization. SODA successively optimizes the Fisher score for each column vector in the matrix.

Formulation (SODA)**.** *Matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_k]$ defines a map $\mathbb{R}^m \to \mathbb{R}^k$, whose columns satisfy:*

$$\begin{aligned} \underset{\mathbf{w}_i}{\textit{maximize}} \quad & \frac{\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i}{\mathbf{w}_i^T\mathbf{S}_W\mathbf{w}_i} \\ \textit{subject to:} \quad & \mathbf{w}_i \perp \mathbf{w}_{j\neq i}, \mathbf{w}_i^T\mathbf{w}_i = 1, \mathbf{w}_i \in \textit{Span}(\mathbf{S}_W) \end{aligned} \tag{11}$$

where $\textit{Span}(\mathbf{S}_W)$ denotes the range space of matrix $\mathbf{S}_W$. □

The implementation can be found in [5]. For binary classification problems, no eigenvectors need to be computed, but only vector operations are involved.

3.2.2. Kernel SODA

The kernel counterpart of SODA has been presented in [23], which is recapped as follows

Formulation (KSODA)**.** *Find optimal vectors $\mathbf{a}_1 \cdots \mathbf{a}_k$, such that:*

$$\begin{aligned} \underset{\mathbf{a}_i}{\textit{maximize}} \quad & \frac{\mathbf{a}_i^T\mathbf{M}\mathbf{a}_i}{\mathbf{a_i}^T\mathbf{N}\mathbf{a}_i} \\ \textit{subject to} \quad & \mathbf{a}_i^T\mathbf{K}\mathbf{a}_{j\neq i} = 0, \mathbf{a}_i^T\mathbf{K}\mathbf{a}_i = 1, \mathbf{a}_i^T\mathbf{N}\mathbf{a} > 0 \end{aligned} \tag{12}$$

Techniques in SODA family provide a good approximation to the SoSNR optimization problems without requiring the nonlinear optimization procedures.

3.3. Sparsity Based Algorithm

In a feature vector, each feature will incur certain cost for its acquisition. This motivates us to impose sparsity constraints. The $l1$ regularization can be extended to matrix case, which is specified as below.

Formulation (Sparse SoSNR Optimization)**.** *Given training data, k and λ, find matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k]$, such that*

$$\begin{aligned} \underset{\mathbf{w}_1,\cdots,\mathbf{w}_k}{\textit{maximize}} \quad & \sum_{i=1}^{k} \frac{\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i}{\mathbf{w}_i^T\mathbf{S}_W\mathbf{w}_i} \\ \textit{subject to:} \quad & \|\mathbf{w}_i\|_1 \leq \lambda,\ \ \mathbf{w}_i^T\mathbf{w}_j = \delta_{ij}. \end{aligned} \tag{13}$$

where δ_{ij} is the Kronecker delta function.

In the subsequent simulation studies, interior-point optimization technique is adopted. As a result, by imposing sparsity, feature selection and transformation are performed simultaneously.

4. EXPERIMENTAL RESULTS

4.1. Data Description

Experiments are based on data sets from the UCI data base [24]. There are four datasets, wdbc, Sonar, Climate modeling and vehicle classification. Among which, vehicle is a multi-class data set with four classes, whereas in this paper we are mainly tackling feature reduction issue for binary classification. Thus vehicle is treated as four binary classification problems in a one-against-all fashion. All features are normalized with respect to their maximum value.

4.2. Parameter and Test

Parameters (such as kernel configurations) are chosen by cross-validation. Tests are performed by 10-fold validation on each data set.

4.3. Experiments and Results

Experiments are designed to verify the following hypotheses:

- The strong correlation between the proposed criterion SoSNR and classification performance, i.e. a higher SoSNR implies that it is highly probable that by using such features it will result in a better classification performance (cf. Figure 3, 4). The classification model used is linear SVM [21].

Data set		Dimension	Original	PCA	KPCA	SODA	KSODA	SO	KSO
wdbc		$30 \rightarrow 8$	7.86%	9.10 %	6.10%	5.05%	5.42%	5.05 %	5.56%
sonar		$60 \rightarrow 8$	26.60%	27.22 %	30.88 %	24.788%	11.32%	23.94%	11.17 %
Climate[25]		$18 \rightarrow 8$	13.73%	28.10%	27.04%	14.22%	12.02%	14.08 %	13.21%
vehicle	van	$18 \rightarrow 8$	3.27%	4.79 %	9.17%	4.16%	3.21%	3.75%	3.26 %
	saab	$18 \rightarrow 8$	12.64%	8.52 %	18.75%	7.92%	7.41%	8.02%	7.09%
	bus	$18 \rightarrow 8$	10.99%	9.18%	12.16%	7.39%	6.59%	7.01 %	6.40%
	opel	$18 \rightarrow 8$	9.88%	6.88 %	12.81 %	6.13%	5.71%	5.96%	5.34 %

Table 1. Classification error comparison between different feature reduction techniques.

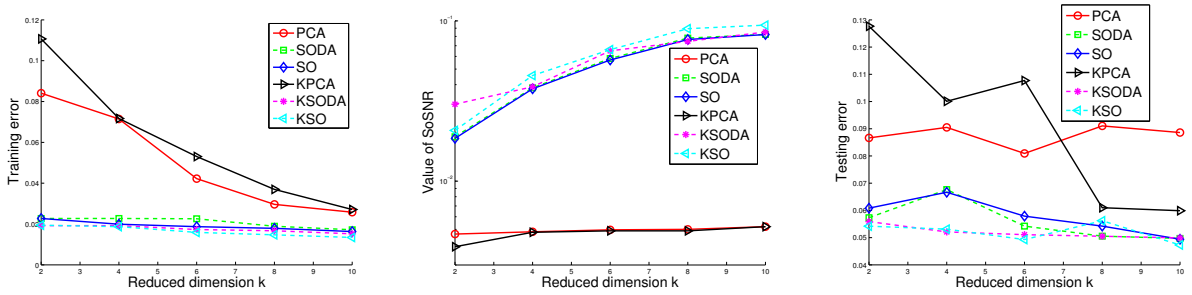

Fig. 3. Training and testing classification error rate and the corresponding SoSNR value of training data on UCI data set WDBC using different dimensionality reduction scheme.

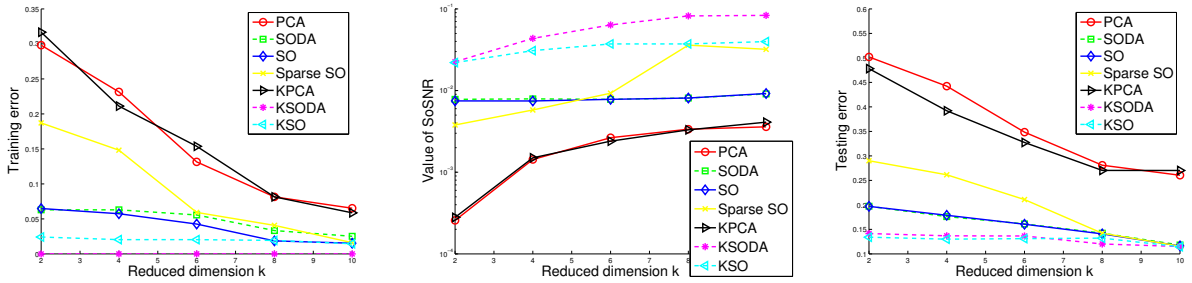

Fig. 4. Same comparison as in Figure 3 on UCI data set CLIMATE. In this figure, we also include the results from Sparse SoSNR Optimization model. Focus of attention is placed on between 8 and 10 features, since they can produce good classification results with reasonable computational complexity.

- Comparison, in terms of (a) classification errors and (b) SoSNR values, between various feature reduction techniques including: I) PCA/KPCA [7, 12], II) SODA/KSODA and III) SoSNR Optimizer (SO)/Kernel SO (KSO) using interior-point method with 50 Monte Carlo simulations (cf. Figure 3, 4 and Table 1). In the presented examples, as unsupervised techniques, PCA/KPCA do not take into consideration of "teacher's information" and hence do not provide the best information regarding class sparability. Moreover, SoSNR Optimization (SO) outperforms SODA in most of the cases, but with higher computational complexity especially when k is large. Nevertheless, we recommend SO over SODA for small k and vice versa.
- Classification error comparison with or without sparsity constraint for SoSNR Optimization (cf. Figure 4, 5). As we can see, when k is very small, Sparse SO is not among the best choices, since a sparse representation of the transformation matrix might discard some significant features and therefore even more information is threw away compared to the other techniques. However, with an increasing k, more combinations of the features are included, which makes sparse model a valid choice. The example in Figure 5 shows the obtained $\mathbf{W}^*$ based on the same testing classification accuracy on the data set sonar. The simulation suggests that on average the sparsity achieved is around 65% ($\lambda = 2.5$).
- Performance with respect to k value (cf. Figure 3 4).

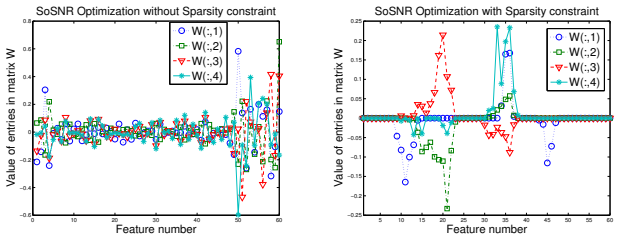

Fig. 5. Obtained $\mathbf{W}$ with and without sparsity constraint.

4.4. Future work

More empirical tests and theoretical results are under progress, including: (1) global optimality and convergence analysis; (2) theoretical results concerning the relation between SoSNR and classification performance for various types of classifiers; (3) theoretical evaluations of Algorithm SODA/KSODA as an approximation of SO; (4) alternative formulations and implementations for Sparse SO with more testing results and (5) other choices of representations for SNR.

5. REFERENCES

[1] Schaible S. and Shi J., *Fractional programming: The sum-of-ratios case.* Optimization Methods and Software, Vol. 18, Issue 2, pp. 219-229, 2003.

[2] Fukumizu, K., Bach, F. R. and Jordan, M. I., *Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces.* The Journal of Machine Learning Research, Vol. 5, pp. 73-99, Jan. 2004.

[3] Hastie, T., Tibshirani, R., and Friedman J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition, Springer, Feb. 2009

[4] Guyon I. and Elisseeff A., *An introduction to variable and feature selection.* The Journal of Machine Learning Research, vol. 3, pp. 1157-1182, March 2003.

[5] Yu Y., Mckelvey T. and Kung S.Y., *A Classification Scheme for 'High-Dimensional-Small-Sample-Size' Data Using SODA and Ridge-SVM with Microwave Measurement Applications.* Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013.

[6] Duda, R.O., Hart, P.E. and Stork, D.G., *Pattern Classification*, 2nd Edition, John Wiley & Sons, New York, 2011.

[7] Jolliffe I.T., *Principal Component Analysis.* Series: Springer Series in Statistics, 2nd ed., Springer, 2002.

[8] Frieden, B. R. *Science from Fisher Information: A Unification*, Cambridge Univ. Press, 2004.

[9] Stephen B.; Lieven V., *Convex Optimization* Cambridge University Press, 2004.

[10] Quanquan Gu, Zhenhui Li, Jiawei Han, *Generalized Fisher Score for Feature Selection.* The 27th Conference on Uncertainty in Artificial Intelligence (UAI), Barcelona, Spain, 2011.

[11] Hoerl A. E. and Kennard R. W. , *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, vol. 12, No. 1, pp. 55-67 Feb., 1970.

[12] Diamantaras K. I., Kung S.Y., *Principal component neural networks: theory and applications.* John Wiley & Sons, 1996.

[13] Mika S., Ratsch G., Weston J., Scholkopf B., and Mullers K. R., *Fisher discriminant analysis with kernels.* Proceedings of the IEEE Signal Processing Society Workshop in Neural Networks for Signal Processing IX, pp. 41 - 48, Aug 1999

[14] Slavakis K., Bouboulis P., Theodoridis S., *Online Learning in Reproducing Kernel Spaces.* E-reference for Signal Processing, Elsevier, 2013.

[15] Kung S.Y., *Kernel Methods and Machine Learning.* Cambridge University Press, 2014.

[16] Fisher R. A., *The use of multiple measurements in taxonomic problems*, Annals of Eugenics, vol. 7, pp. 179-188, 1936.

[17] Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of Royal Statistics Society: Series B, Vol. 58, No. 1, pages 267-288, 1996.

[18] Ye, J, Xiong, T, *Computational and Theoretical Analysis of Null Space and Orthogonal Linear Discriminant Analysis.* Journal of Machine Learning Research vol. 7, pp. 11831204, 2006.

[19] Ng, A. Y., *Feature selection, L1 vs. L2 regularization, and rotational invariance.* 21st ICML, New York, USA, ACM, 2004.

[20] Li, F., Yang, Y., Xing, E. P., *From Lasso regression to Feature vector machine.* NIPS. 2003.

[21] Vapnik, Vladimir N., *Statistical Learning Theory*, 1st Edition, Wiley-Interscience, September, 1998.

[22] Daniela M. Witten, Robert Tibshirani, *Penalized classification using Fisher's linear discriminant.* Journal of the Royal Statistical Society: Series B, Vol. 73, No. 5, pp. 753772, Nov 2011.

[23] Yu Y., McKelvey T., and Kung S.Y, *Kernel SODA: A Feature Reduction Technique Using Kernel Based Analysis*, The 12th International Conference on Machine Learning and Applications (ICMLA'13), Miami, Florida, USA, Dec. 4 - 7, 2013.

[24] http://archive.ics.uci.edu/ml/.

[25] Lucas D. D., Klein R., Tannahill J., Ivanova D., Brandon S., Domyancic D., and Zhang, Y., *Failure analysis of parameter-induced simulation crashes in climate models.* Geosci. Model Dev. Discuss., 6, 585-623, 2013.