A UNIFIED SUBSPACE CLASSIFICATION FRAMEWORK DEVELOPED FOR DIAGNOSTIC SYSTEM USING MICROWAVE SIGNAL

Yinan Yu, Tomas McKelvey

Chalmers University of Technology, Gothenburg, Sweden

ABSTRACT

Subspace learning is widely used in many signal processing and statistical learning problems where the signal is assumably generated from a low dimensional space. In this paper, we present a unified classifier including several concepts from different subspace techniques, such as PCA, LRC, LDA, GLRT, etc. The objective is to project the original signal (usually of high dimension) into a smaller subspace with 1) within-class data structure preserved and 2) between-class-distance enhanced. A novel classification technique called Maximum Angle Subspace Classifier (MASC) is presented to achieve these purposes. To compensate for the computational complexity and non-convexity of MASC, an approximation is proposed as a trade-off between the classification performance and the computational issue. The approaches are applied to the problem of classifying high dimensional frequency measurements from a microwave based diagnostic system and results are compared with existing methods.

***Index Terms*—** Supervised subspace learning, classification, high dimensional data, class separability

1. INTRODUCTION

Over the last decades, a family of techniques called Matched Subspace Detector (MSD) were proposed for signal detection problems [1, 2]. These techniques are based on subspace models and statistical assumptions of the signal. On the other hand, in the machine learning society, there have been also a fair number of papers on similar techniques dealing with classification problems[3, 4, 5, 6]. The underlying idea is that samples from each class are generated from an individual subspace. In paper [4], a technique called Linear Regression Classifier (LRC) is introduced. Basically, under the subspace assumption, LRC projects the testing data onto the span of the training data from each class and identifies the subspace with smallest projection error. There exist many extended versions of LRC, which mostly focus on improving the estimation of the subspaces by reducing the regression error using the training data.

However, in a classification problem, the between-class distance plays a key role as a part of the class separability measure and a proper metric needs to be defined in such cases. In this work, we present a unified subspace classification technique called Inner-product Subspace Classifier (ISC) and there are two steps involved: 1) constructing individual basis for each class, and 2) computing the inner product of the projected testing vector.

Under the ISC framework, a Naive ISC is first presented as the simplest classification model, where the basis for each subspace is estimated using Singular Value Decomposition (SVD) from the training data. In Section 3.2, a novel technique called Maximum Angle Subspace Classifier (MASC) is proposed. MASC explicitly maximizes the between-class distance while keeping the within-class variation small. However, it suffers from a high computational complexity and no global optimum is guaranteed. Motivated by these facts, a new method called Empirical Subspace Intersection Removed Classifier (ESIRC) is presented as a trade-off between the classification performance and computational complexity in Section 3.3. By removing nearly common directions, this technique enhances the class separability without consuming more computational power.

The developed classifiers are applied to data measured by a newly developed near field microwave system consisting of transceivers connected to an array of antennas placed around the object under study. The objective of this system is to detect objects with an internal anomaly. Empirical results show that MASC gives the best classification rate among the compared techniques. ESIRC serves as an efficient approximation and provides a good trade-off between computational complexity and classification accuracy.

2. A SHORT REVIEW OF RELATED WORK

The technique most closely related to ISC is Linear Regression Classifier (LRC) [4], which has been proposed for face recognition in 2010. In this technique, the classification criterion is to find the minimum projection error of the testing data to all the subspaces. Several modified approaches have been presented as well, such as Principal Component Regression Classifier (PCRC), Improved Principal Component Regression Classifier (IPCRC) [5], Ridge Regression Classifier (RRC) [6], etc. PCRC finds the regression coefficients which minimize the residual errors in PCA space, IPCRC discards the first principal components to improve the robustness of the classifier in a varying environment and RRC is proposed to tackle degenerated cases.

In this paper, a technique called Naive ISC is introduced. Algebraically, Naive ISC is equivalent to LRC. The only difference in the formulation is that instead of finding the smallest projection error, Naive ISC identifies the largest subspace inner product. The reason is that LRC focuses on the 'goodness of fitting' in a regression sense, and hence the criterion is naturally based on the size of the regression residual.

Moreover, compared to LRC related techniques, extensions of ISC are different in the following ways: 1) instead of manipulating the subspace coordinates, the whole subspace for each class is operated as a point on the Grassmann Manifold, where metrics are naturally defined. The basis matrices are then considered as variables for optimization and constraints can be defined on a subspace level. 2) Both between-class and within-class characteristics are taken into consideration simultaneously. As suggested by the name, LRC and its extensions emphasize on minimizing the regression error which guarantees a small within-class scattering. However, for classification problems, the between-class distance has a great contribution to the class separability. In ISC, between-class distance is defined using the metric equipped by the corresponding Grassmannian. 3) Inner product space is defined for individual subspace and kernel tricks can be applied to improve the classification accuracy [7].

3. A UNIFIED FRAMEWORK

Given C, the total number of classes, let $\boldsymbol{x}_c^i$ be a sample drawn from class $c \in \{1, \cdots, C\}$. We assume that $\boldsymbol{x}_c^i$ is generated according to the following data model:

$$\boldsymbol{x}_c^i = \sum_{k=1}^{N_c} \boldsymbol{u}_{c,k}^0 \alpha_c^i(k) + \boldsymbol{e}_c^i = \boldsymbol{U}_c^0 \boldsymbol{\alpha}_c^i + \boldsymbol{e}_c^i \tag{1}$$

where $\boldsymbol{U}_c^0$ is a matrix containing the basis vectors $\boldsymbol{u}_{c,k}^0$ as its columns and $\boldsymbol{\alpha}_c^i$ represents the corresponding basis weight vector. The vector $\boldsymbol{e}_c^i$ is the error between the model $\sum_{k=1}^{N_c} \boldsymbol{u}_{c,k}^0 \alpha_c^i(k)$ and the measurement $\boldsymbol{x}_c^i$. Without ambiguity, $\boldsymbol{U}_c^0$ is used to denote the linear subspace from now on. Note that given the feature dimension p, we assume that $N_c \ll p$ always holds.

For a given class c, the columns of $\boldsymbol{U}_c$ is the estimated orthonormal basis vectors spanning the subspace. Hence, the inner product of a testing sample $\boldsymbol{x}$ in subspace $\boldsymbol{U}_c$ is defined as:

$$\begin{aligned} l_c^2(\boldsymbol{x}) &= \|\boldsymbol{P}_c \boldsymbol{x}\|_2^2 \\ &= \boldsymbol{x}^H \boldsymbol{U}_c \boldsymbol{U}_c^H \boldsymbol{x} \end{aligned} \tag{2}$$

where, $\boldsymbol{P}_c = \boldsymbol{U}_c \boldsymbol{U}_c^H$ is the projection matrix which projects $\boldsymbol{x}$ onto the $\boldsymbol{U}_c$ space. Class membership of $\boldsymbol{x}$ is determined either as the largest size within the class

$$\hat{c} = \arg\max_c l_c(\boldsymbol{x}) \tag{3}$$

This is the ML classification for suitable statistical assumptions. Equivalently, we can also use the shortest distance to the subspace, which coincides with the criterion for LRC [4].

$$\hat{c} = \arg\min_c \|\boldsymbol{x} - \boldsymbol{P}_c \boldsymbol{x}\|_2 \tag{4}$$

In summary, Inner-product Subspace Classifier (ISC) is formally formulated in Definition 1. Note that without loss of generality, we restrict the presentation to the case of a binary classifier, i.e. $c \in \{1, 2\}$. The generalization to multi-class problems can be viewed as a natural extension of the binary case.

Definition 1. *(Binary) Inner-product Subspace Classifier (ISC)*

Given basis $\boldsymbol{U}_1$, $\boldsymbol{U}_2$ and testing data $\boldsymbol{x}$, let:

$$\delta(\boldsymbol{x}) = \boldsymbol{x}^H \boldsymbol{U}_1 \boldsymbol{U}_1^H \boldsymbol{x} - \boldsymbol{x}^H \boldsymbol{U}_2 \boldsymbol{U}_2^H \boldsymbol{x} \tag{5}$$

ISC is defined as a function $f(\boldsymbol{x})$, such that:

$$f(\boldsymbol{x}) = \begin{cases} +1 & \text{if } \delta(\boldsymbol{x}) > 0 \\ -1 & \text{if } \delta(\boldsymbol{x}) \leq 0 \end{cases} \tag{6}$$

□

In a more general setting, instead of only considering the criterion introduced in (3), a classifier such as Support Vector Machine (SVM) [8] can be applied to the two dimensional space containing points $\{(l_1^2, l_2^2)\}$ to obtain the predicted labels. Now the question is how can we construct $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ to improve the performance of the classifier.

3.1. The simplest model

A simplest mode called Naive ISC is defined as an ISC equipped with a straightforward estimation of the projection matrix. That is, in the training step, we construct a data matrix $\boldsymbol{X}_c$ by placing all the training data from class c as its columns:

$$\boldsymbol{X}_c = \left[\boldsymbol{x}_c^1,\ \boldsymbol{x}_c^2,\ \cdots,\ \boldsymbol{x}_c^{m_c}\right] \tag{7}$$

In our study, we assume that the training sample size m_c is always much smaller than the feature dimension p and $N_c = m_c$. Therefore, given $\boldsymbol{X}_c$, the basis $\boldsymbol{U}_c$ can be estimated by QR or singular value decomposition of $\boldsymbol{X}_c$.

The projection matrix $\boldsymbol{P}_c$ presented in Equation (2) can be estimated by:

$$\boldsymbol{P}_c = \boldsymbol{X}_c (\boldsymbol{X}_c^H \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^H \tag{8}$$

Without any modification, Naive ISC is equivalent to LRC and closely related to the Generalized Likelihood Ratio Test (GLRT) [9], which is included as a special case of ISC.

3.2. An improved model with high complexity

We are interested in finding subspaces $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ such that the class separability is enhanced. As we know, subspaces can be considered as points on a Grassmann manifold $G(k, n)$ and the corresponding metrics can be defined to study the topology of the set. With a well defined searching space using these terminologies and their corresponding characteristics, optimization can be applied to locate candidates of the subspace basis.

First, we define Grassmann manifold and its equipped metric in order to formally describe the searching space of MASC.

Definition 2. *Grassmann manifold[13]*

Let V be a complex vector space of dimension n. The Grassmann Manifold, denoted by $G(k, V)$ is defined to be the set of k-dimensional linear subspaces of V; we write $G(k, n)$ for $G(k, \mathbb{C}^n)$. □

Therefore, a subspace constructed by training data from each class is one point on the Grassmann manifold. Metrics can be equipped by Grassmann manifold in order to assess the between-classes distance. One widely used metric is the principal angles, which is defined as follows:

Definition 3. *Principal angles [14]*

The principal angles θ_k between the subspaces $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are defined as:

$$\begin{aligned} \cos(\theta_k) &= \max_{\boldsymbol{u} \in \boldsymbol{U}_1} \max_{\boldsymbol{v} \in \boldsymbol{U}_2} \boldsymbol{u}^H \boldsymbol{v} = \boldsymbol{u}_k^H \boldsymbol{v}_k \\ & \textit{subject to:} \\ & \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1 \\ & \boldsymbol{u}^H \boldsymbol{u}_i = 0, \quad i = 1, \cdots, k-1 \\ & \boldsymbol{v}^H \boldsymbol{v}_i = 0, \quad i = 1, \cdots, k-1 \end{aligned} \tag{9}$$

Empirical results show that with a constraint on the projection error of the training data, when the principal angles are larger, the classification performance becomes better. This is due to the fact that compared to subspaces with small angles, for the same level of noise, the data points are less likely to be mixed. This motivates us to find subspaces with a maximum minimal angle between them without sacrificing the accuracy of the subspace representation of the data. The Maximum Angle Subspace Classifier (MASC) defines such a basis:

Definition 4. MASC

A Maximum Angle Subspace Classifier (MASC) with accuracy ϵ is defined by the bases $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ such that:

$$\begin{aligned} &\underset{\boldsymbol{U}_1,\boldsymbol{U}_2}{\text{minimize}} && \max_{\boldsymbol{u}\in\boldsymbol{U}_1}\max_{\boldsymbol{v}\in\boldsymbol{U}_2} \boldsymbol{u}^H\boldsymbol{v} \\ &\text{subject to} && \frac{1}{m_c}\sum_{i=1}^{m_c}(\|\boldsymbol{x}_c^i - \boldsymbol{U}_c\boldsymbol{U}_c^H\boldsymbol{x}_c^i\|_{l_2}^2) \leq \epsilon \\ &&& \boldsymbol{U}_c^H\boldsymbol{U}_c = \mathbf{I}, \end{aligned} \tag{10}$$

where m_c is the number of samples in class c and $\boldsymbol{U}_c \in G(N_c, p)$. □

A heuristic algorithm which sequentially maximizes the minimal angle can be found in Algorithm 1. The notation '$\sim c$' is used to denote the classes 'not c'. The optimization problem in Algorithm 1 is approached numerically using a constrained nonlinear programming algorithm. Note that due to the high computational complexity, a subspace projection step (PCA) needs to be applied before Algorithm 1 for dimensionality reduction.

Algorithm 1 Construction of MASC basis

- Apply PCA for dimensionality reduction
- Initialize $\boldsymbol{U}_c$ by QR or SVD of the data matrix
- For $c \in \{1, ..., C\}$, iterate until convergence:

Fix $\boldsymbol{U}_{\sim c}$:

$$\begin{aligned} &\underset{\boldsymbol{U}_c\in G(N_c,p)}{\text{minimize}} && \max_{\boldsymbol{u}\in\boldsymbol{U}_c}\max_{\boldsymbol{v}\in\boldsymbol{U}_{\sim c}} \boldsymbol{u}^H\boldsymbol{v} \\ &\text{subject to} && \frac{1}{m_c}\sum_{i=1}^{m_c}(\|\boldsymbol{x}_c^i - \boldsymbol{U}_c\boldsymbol{U}_c^H\boldsymbol{x}_c^i\|_{l_2}^2) \leq \epsilon \\ &&& \boldsymbol{U}_c^H\boldsymbol{U}_c = \mathbf{I} \end{aligned} \tag{11}$$

3.3. A complexity reduced approximation

MASC maximizes the between-class distance while keeping the within-class scattering small. However, it is relatively time consuming and no global optimum can be guaranteed. In this section, a different approach is attempted. Instead of directly maximizing the principal angles, the bases are reduced by removing the direction(s) in respective basis which have the smallest principal angle(s) with respect to the other basis. This improves the classification performance when the bases are estimated from noisy data.

If some of the principal angles between the spaces would be zero the two spaces would have a common intersecting subspace. However, in practice the angles are never zero and we call the directions which have angles close to zero as the Empirical Subspace Intersection (ESI) for each subspace respectively. Formaly we define ESI as follows:

Definition 5. Empirical Subspace Intersection

The Empirical Subspace Intersection $ESI(\boldsymbol{U}_1, \boldsymbol{U}_2, \delta)$ of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ is defined as:

$$ESI(\boldsymbol{U}_1, \boldsymbol{U}_2, \delta) = \{\boldsymbol{u}_j : \boldsymbol{u}_j \in \boldsymbol{U}_1,\ \forall\, j,\ s.t.\ \theta_j < \delta\} \tag{12}$$

where s is the dimension of the empirical subspace intersection, θ_j's are the principal angles and $\boldsymbol{u}_j$ are the the principal vectors and δ is the empirical tolerance. □

The algorithm of removing $ESI(\boldsymbol{U}_1, \boldsymbol{U}_2, \delta)$ from $\boldsymbol{U}_1$ and $ESI(\boldsymbol{U}_2, \boldsymbol{U}_1, \delta)$ from $\boldsymbol{U}_2$ is summarized in Alg.2 [14].

Algorithm 2 ESI removal between 2 subspaces:

Note: in this algorithm, the intermediate basis is called $\boldsymbol{Q}_i$, $i = 1, 2$. The notation $\boldsymbol{U}_i$, $i = 1, 2$ is used to denote the final constructed basis for class i.
- Let the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be vectors spanning subspace 1 and 2 respectively and $N_1 \geq N_2$ be the dimensions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ respectively;
- Compute the QR:

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{Q}_1\boldsymbol{R}_1 \\ \boldsymbol{X}_2 &= \boldsymbol{Q}_2\boldsymbol{R}_2 \end{aligned} \tag{13}$$

- Construct matrix $\boldsymbol{C}$: $\boldsymbol{C} = \boldsymbol{Q}_1^T\boldsymbol{Q}_2$
- Compute the SVD of $\boldsymbol{C}$ (assume $N_1 > N_2$):

$$Y^T\boldsymbol{C}Z = \begin{bmatrix} \cos(\theta_1) & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & \cos(\theta_{N_2}) \\ 0 & \vdots & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{14}$$

where θ_k's are the principal angles and $1 \geq \cos(\theta_1) \geq \cos(\theta_2) \geq \cdots \geq \cos(\theta_s) >> \cos(\theta_{s+1})$.
- Compute the associated basis $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_k\}$:

$$\begin{aligned} \boldsymbol{Q}_1\boldsymbol{Y} &= [\boldsymbol{u}_k], k = 1, \cdots, N_1 \\ \boldsymbol{Q}_2\boldsymbol{Z} &= [\boldsymbol{v}_k], k = 1, \cdots, N_2 \end{aligned} \tag{15}$$

where

$$\begin{aligned} ESI(\boldsymbol{Q}_1, \boldsymbol{Q}_2, \theta_{s+1}) &= [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \boldsymbol{u}_s] \\ ESI(\boldsymbol{Q}_2, \boldsymbol{Q}_1, \theta_{s+1}) &= [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots \boldsymbol{v}_s] \end{aligned} \tag{16}$$

- Remove the intersection and construct the subspaces:

$$\begin{aligned} \boldsymbol{U}_1 &= [\boldsymbol{u}_{s+1}, \boldsymbol{u}_{s+2}, \cdots, \boldsymbol{u}_{N_1}] \\ \boldsymbol{U}_2 &= [\boldsymbol{v}_{s+1}, \boldsymbol{v}_{s+2}, \cdots, \boldsymbol{v}_{N_2}] \end{aligned} \tag{17}$$

In summary, ESIRC is defined as follows:

Definition 6. ESIRC

An Empirical Subspace Intersection Removal Classifier (ESIRC) is a binary Inner-product Subspace Classifier (ISC) using basis constructed from Algorithm 2. □

4. EXPERIMENTAL RESULTS

The experimental results are based on microwave measurements for the purpose of object classification. There are two prototypes with different measurement setups, which result in various data features summarized in Table 1. The system consists of N antennas mounted around the object under test. Each antenna performs as both transmitter and receiver. The measured signals are scattering parameters, which is defined as the received energy divided by the transmitted energy at each antenna for a single frequency excitation. The signals are measured at 401 different frequencies from 100 MHz to 3.0 GHz. If we vectorize the raw measurement, a high dimensional complex vector space is obtained. The numbers of independent objects from class 1 and 2 are also summarized in Table 1 and for each object, multiple measurements are taken.

In our experiment, subspace basis for each class is learned using: (a) singular value decomposition of the data matrix defined in Equation (7) for Naive ISC (b) empirical intersection removal by Algorithm 2 and (c) maximization of principal angles using Algorithm 1 (MASC). Note that for MASC, feature reduction using PCA is performed on the dataset. The new dimension of the feature space is $m_1 + m_2$. On the other hand, no dimension reduction is needed for ISC and ESIRC. The experimental results are compared with other supervised classification approaches: (d) linear SVM (e) Gaussian kernel SVM [15], (f) PCA+ LDA, (g) PCRC [5]. and (h) IPCRC [5].

	Dataset 01	Dataset 02
Number of antennas: N	10	12
Dimension of the signal	22,055	31,278
Measurements in class 1	27	30
Measurements in class 2	21	45
Independent objects in class 1	11	15
Independent objects in class 2	9	11

Table 1. Description of settings for dataset 01 and dataset 02.

All classifiers are evaluated using a leave-one-out validation procedure [16]. That is, for each run, all measurements from one object is removed for testing. For training, one measurement is randomly picked from each of the remaining objects. The empirical performance of the classifiers are recorded for all the left-out measurements and shown in Table 2. The given evaluation is the classification accuracy of class 1 at a constant false alarm rate of 20%. In our experiments, PCA + LDA technique improves the classification results compared to linear SVM, whereas by applying Gaussian kernel to SVM, it provides a more significant improvement. Probably due to the large dimensionality and small sample size issue, PCRC does not increase the classification accuracy. This is because when the sample size is very small, the estimated basis does not converge to the true subspace and thus even components corresponding to small eigenvalues contain information which makes the PCA space less representative. Nevertheless, IPCRC gives a better result compared to LRC and PCRC in our case. The reason is that for medical measurements, the varying environments and situations might dominate the main variability. By discarding the first principal components, these effects are excluded to some extent. MASC searches for two representative subspace bases with maximal principal angles on Grassman manifold and provides the best performance among all the compared techniques on both datasets. EESIRC ended up being a good trade-off.

To further compare the three proposed classification schemes: ISC, ESIRC and MASC, scatter plots of the inner products (l_2^2, l_1^2) are shown in Figure 1 for dataset 01 and 02 respectively. The only difference of the three methods is that they use different basis construction techniques. As we can see from the figure, MASC results in the best class separability. However, if computational time is of concern, ESIRC can be a good substitute.

	Classification model	Dataset 01	Dataset 02
(a)	LRC (ISC)	57%	56%
(b)	ESIRC	85%	76 %
(c)	MASC	92%	89 %
(d)	Linear SVM	57%	55 %
(e)	(Gaussian) Kernel SVM	78 %	73 %
(f)	PCA+LDA	67 %	64%
(g)	PCRC	57%	54%
(h)	IPCRC	77 %	72%

Table 2. Empirical performance of different classifiers for class 1 at a constant false alarm rate of 20%. The parameters are chosen as: (a) $N_c = m_c$; (b) $N_c = m_c -$ dim(ESI), where dim(ESI)= 4 for data 01 and dim(ESI)= 6 for data 02; (c) $N_c = m_c$; (d) $C = 10$; (e) Gaussian kernel $\sigma = 1$, $C = 10$; (f) dimension of PCA space is m_c; (g,h) $N_c = m_c - k_c$, where $k_c = 4$ for data 01 and $k_c = 6$ for data 02.

5. CONCLUSION

In this paper, a novel subspace classification framework called ISC is proposed. ISC includes several subspace detection and classification techniques as special cases and it actively searches the optimal subspaces on a Grassman manifold. The underlying idea of ISC is to enhance the class separability through different subspace constructions. Compared to other class separability maximization techniques, such as LDA, the classifiers proposed in this paper are able to deal with subspace data model, in which case, LDA does not give the optimal solution. From empirical tests, constructed MASC gives the best classification result among all the compared techniques. However, when the dimension of the subspace grows, the optimization becomes highly time consuming. Another drawback of this approach is that there is no global optimum guaranteed. On the other hand, ESIRC provides a good tradeoff between computational complexity and classification performance. Our future work will focus on 1) estimation of the subspace dimensionality and 2) further studies of how the parameters in the proposed algorithm affect the classification performance.

6. ACKNOWLEDGMENT

This work was supported by the Swedish Research Council (VR) which is gratefully acknowledged.

7. REFERENCES

[1] Scharf L. and Friedlander B., *Matched subspace detectors*, IEEE Transactions on Signal Processing (TSP), Vol. 42, Issue 8, pp: 2146-2157, 1994.

[2] Besson O. and Scharf L., *CFAR matched direction detector*, IEEE Transactions on Signal Processing (TSP), Vol. 54, Issue 7, pp: 2840-2844, 2006.

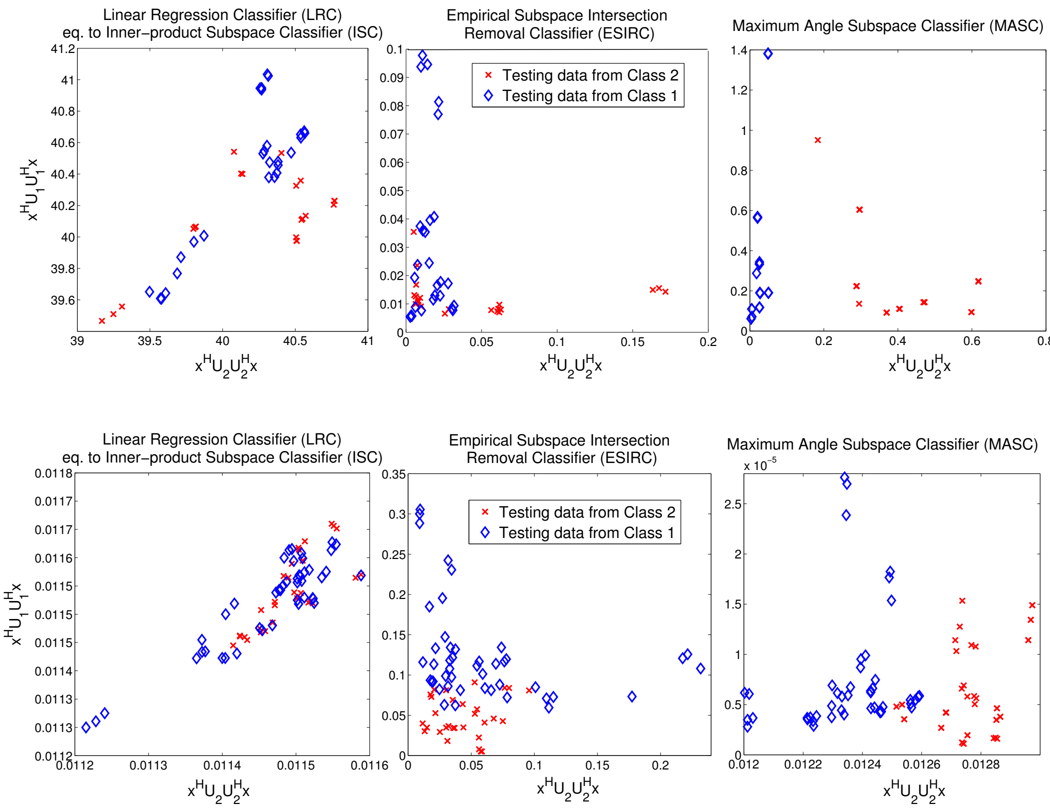

Fig. 1. The figures have shown the inner product (l_2^2, l_1^2) of test samples from dataset 01 (upper) and 02 (lower) in both subspaces using different basis constructions. Red crosses and blue diamonds correspond to samples from Class 1 and Class 2 respectively. [Left] Naive ISC, [middle] ESIRC and [right] MASC. As we can see, the scatter plot resulting from MASC is more clearly separated compared to ISC and ESIRC. However, ESIRC serves as a trade-off between the classification accuracy and computational complexity.

[3] Yan S., Liu J., Tang X, and Huang T., *A Parameter-Free Framework for General Supervised Subspace Learning*. Information Foresniscs and security, Volume 2, Issue 1 pp: 69 - 76 , 2007.

[4] Naseem I., Togneri R., and Bennamoun M., *Linear Regression for Face Recognition*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 11, pp. 2106-2112, July 2010.

[5] Huang S. and Yang J., *Improved Principal Component Regression for Face Recognition under Illumination Variations*, IEEE Signal Process. Lett., vol. 19, no. 4, pp. 179-182, April 2012.

[6] Arpit D., Wu S., Natarajan P., Prasad R. and Natarajan P., *Ridge Regression based classifiers for large scale class imbalanced datasets*, Applications of Computer Vision (WACV), 2013 IEEE Workshop on, pp. 267 - 274, 15-17 Jan. 2013.

[7] Scholkopf B, Smola A. and Muller K., *Kernel principal component analysis*, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, pp. 327–352, MIT Press, 1999.

[8] Chapelle O., Haffner P., and Vapnik V., *Support Vector Machines for Histogram-Based Image Classification*, IEEE Transactions on Neural Networks, VOL. 10, NO. 5, Sep. 1999.

[9] Kay S., *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, Prentice-Hall, 1998.

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

[11] Van Trees H., *Detection, Estimation, and Modulation Theory*, New York: Wiley, 1968.

[12] Strang, G., *Introduction to Linear Algebra*, Wellesley-Cambridge Press and SIAM, 2009.

[13] Absil P.-A., Mahony R. and Sepulchre R., *Riemannian geometry of Grassmann manifolds with a view on algorithmic computation*, Acta Appl. Math, January, 2004.

[14] Golub, G. and Van Loan C. F., *Matrix Computations*, 3rd edition, Johns Hopkins University Press, 1996.

[15] Kung S.Y., *Kernel Methods and Machine Learning*, Cambridge Press, 2013.

[16] Geisser S., *Predictive Inference*, New York, NY, Chapman and Hall, 1993.