Information and Inference: A Journal of the IMA (2016) Page 1 of 17
doi:10.1093/imaiai/drn000

Consistency and Convergence Rate for Nearest Subspace Classifier

YI (GRACE) WANG,
Department of Mathematics, Duke University, Durham, NC, USA
yiwang@math.duke.edu

[Received on 2 March 2016]

The Nearest subspace classifier (NSS) finds an estimation of the underlying subspace within each class and assigns data points to the class that corresponds to its nearest subspace. This paper mainly studies how well NSS can be generalized to new samples. It is proved that NSS is strongly consistent and has rate of convergence $\mathscr{O}(n^{-1/2})$ under certain assumptions. Some simulations are presented eventually to verify the theoretical results.

Keywords: Nearest Subspace, Classification, Consistency, Rate of Convergence, Supervised Learning

1. Introduction

The problem of classification is to construct a mapping that can correctly predict the classes of new objects, given training examples of old objects with ground truth labels [39]. It is a classical problem in statistical learning and machine learning and has been widely used in computer vision, pattern recognition, bioinformatics, etc. Examples of applications include face recognition, handwriting recognition and micro-array classification.

More precisely, this problem can be formalized as follows. Given a training data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathscr{X}$ and $y_i \in \mathscr{Y}$, the goal is to find a function $f : \mathscr{X} \to \mathscr{Y}$ such that $f(\mathbf{x})$ is a good approximation of y for the given $\mathbf{x}_i$'s as well as for new instances $\mathbf{x}$. Typically, $\mathscr{X}$ is a continuous domain and $\mathscr{Y}$ is a finite discrete set.

In the past few decades, a tremendous amount of work has been produced for this problem. Many approaches have been proposed, e.g., K-Nearest Neighbors (KNN) [17, 21, 24], Fisher's Linear Discriminant Analysis (LDA) [23, 45], Artificial Neural Networks (ANN) [38, 46, 61], Support Vector Machines (SVM) [9, 15, 47], and Decision Trees (see [10, 43, 44] for some well known algorithms). We refer to [6, 27] for a more careful overview of classification techniques.

Among this work is a class of methods based on subspace models. The compelling interest in subspace models can be attributed to their validation in real data. For instance, it has been justified that the set of all images of a Lambertian object (e.g., face images) under a variety of lighting conditions can be accurately approximated by a low-dimensional linear subspace (of dimension at most 9) [5, 22, 28]. Another example is that, under the affine camera model, the coordinate vectors of feature points from a moving rigid object lie in an affine subspace of dimension at most 3 (see [16]). These applications give rise to modeling data by subspaces; the study of subspace based classifiers is an important branch.

The first work in this category was CLAss Featuring Information Compression (CLAFIC) [60] (also known as Nearest SubSpace (NSS) classifier [42]; for the information contained in this name, we will adopt the usage of NSS throughout the paper). In this algorithm, each class is represented by a linear subspace and data instances are assigned to the nearest subspace. Instead of obtaining good representation of subspaces in NSS, the Learning Subspace Method (LSM) [30] proposes to learn the subspaces based on good discrimination (see [40] for more variants and discussions). The simple idea of subspace classifiers has been extended to nonlinear versions in various ways; many have shown state-of-the-art performance (see [12, 34, 53] for example and Section 2.5 for more details). After the first subspace analysis of face images [29, 55], classification approaches with subspace models have been used successfully in face recognition [11], handwritten digit recognition [33], speech recognition [31] as well as biological pattern recognition problems [41].

Although the design of subspace-based classification techniques has been actively explored, their theoretical justification is very under-studied. In this paper, we restrict our interests of justification to analyzing how well the classifiers can be generalized to new samples. By doing so, one can learn quantitatively how reliable the classification approaches are and can thus also guide the algorithm design accordingly. For this purpose, a functional (known as *risk function*) is used to measure the prediction quality of every classifier. More precisely, we assume X and Y being random variables; instances $\mathbf{x}_i$ and y_i are drawn independently from the distributions of X and Y respectively. For a classifier $f(x)$, its risk functional is defined as:

$$R(f) = \mathbb{E}_{(X,Y)} \mathbb{1}(f(X) \neq Y)$$

Based on this, the *Bayes rule* is defined to be the classifier whose risk functional is minimal. The Bayes rule is optimal in the sense that its expected loss (defined as 1 when the predicted class is not equal to the truth) is minimal. Note that, since the actual distribution of (X,Y) is unknown, the Bayes rule is thus not available in reality.

A natural desirable property of practical classifiers is having as small risk functional as possible. In this spirit, the property *consistency* is defined as the fact that the risk function converges to that of the optimal Bayes rule as the number of samples goes to infinity. Many classification algorithms, such as, KNN, SVM, LDA and some boosting methods [1, 4, 8, 50, 52, 57], have been shown to be consistent under certain conditions. Moreover, one would like to learn the convergence rate for a consistent classifier, i.e., how many samples are required to obtain a risk that is close to the optimal risk by a certain small number. This property has been extensively studied for SVM in [3, 7, 14, 49, 51, 58, 62, 63]. The rate of convergence for LDA is investigated in [25] and for KNN is studied in [13, 18, 26, 32] and the references therein.

In this paper, we study the consistency property of the Nearest SubSpace (NSS) classifier. We prove its strong consistency under certain conditions. Furthermore, we study the rate of convergence for the NSS classifier by providing a non-asymptotic bound for the difference between its risk and the optimal risk. This non-asymptotic bound tells how many samples are required to obtain risk that is close to the Bayes risk by a certain small number with overwhelming probability. To our best knowledge, this is the first work on the consistency and convergence rate for the NSS algorithm. Although the techniques used to derive these results have been studied before [19, 36, 37, 57, 66], they have never been applied to the NSS. Our main contribution is to apply these techniques to thoroughly study the NSS classifier and obtain the first result about its consistency and rate of convergence. In the rest of the paper, we will begin with a description of the NSS algorithm and our main theorems (Section 2), followed by their proof (Section 3 and 4) and simulations (Section 5).

2. The NSS Algorithm and our Main Theorems

For most of the applications, it suffices to assume that $\mathscr{X} \subset \mathscr{B}(0,M) \subset \mathbb{R}^D$ and $\mathscr{Y} = \{1,\cdots,K\}$, where $\mathscr{B}(0,M)$ is the ball centered at the origin with radius M and D and K are some positive integers. We will restrict ourselves to this case throughout the paper.

2.1 *The NSS Algorithm*

The NSS classifier assumes data lie on multiple affine subspaces, finds an estimate for these subspaces and assigns each instance to the nearest subspace. The following is a summary of the NSS algorithm.

```
Algorithm 1 Nearest Subspace (NSS) Classification
Require: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathscr{X} \times \mathscr{Y}$ and $d$: intrinsic dimension, some positive integer and $d < D$.
Ensure: A function $f : \mathscr{X} \to \mathscr{Y}$.
  for $k = 1$ to $K$ do
      $$\begin{aligned} \hat{\mathbf{u}}_k &= \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i;\ C_k = \{\mathbf{x}_i : y_i = k\};\ n_k = |C_k|. \\ \hat{\mathbf{B}}_k &= \operatorname*{arg\,min}_{\substack{\mathbf{B} \in \mathbb{R}^{D \times d} \\ \mathbf{B}^T\mathbf{B} = \mathbf{I}_d}} \sum_{\mathbf{x}_i \in C_k} \|(\mathbf{I} - \mathbf{B}\mathbf{B}^T)(\mathbf{x}_i - \hat{\mathbf{u}}_k)\|^2. \end{aligned} \qquad (2.1)$$
  end for
  $\hat{f}(\mathbf{x}) = \operatorname*{arg\,min}_k \|(\mathbf{I} - \hat{\mathbf{B}}_k \hat{\mathbf{B}}_k^T)(\mathbf{x} - \hat{\mathbf{u}}_k)\|^2.$
```

Note that the closed form solution to (2.1) is the Singular Value Decomposition (SVD) of the centered data matrix for the k^{th} class; such a data matrix consists of $\big((\mathbf{x}_{k_1} - \hat{\mathbf{u}}_k), \quad \cdots, \quad (\mathbf{x}_{k_{n_k}} - \hat{\mathbf{u}}_k)\big)$ with $\mathbf{x}_{k_1}, \cdots, \mathbf{x}_{k_{n_k}} \in C_k$.

2.2 *Notations*

Denote L_k as the underlying d-dimensional subspace for the k^{th} class; denote $\mathbf{u}_k$ as the underlying center and $\mathbf{B}_k$ as an underlying orthonormal basis for the k^{th} class. Let $\mathbf{P}_k = \mathbf{B}_k\mathbf{B}_k^T$ and $\hat{\mathbf{P}}_k = \hat{\mathbf{B}}_k\hat{\mathbf{B}}_k^T$. Denote $\hat{L}_k$ as the subspace spanned by $\hat{\mathbf{B}}_k$. Assume that $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ are i.i.d. samples of random variable (X, Y); $X \in \mathbb{R}^D$ and $Y \in \{1, \ldots, K\}$. Then let $\Sigma_k = \mathbb{E}_{X \in C_k}(X - \mathbf{u}_k)(X - \mathbf{u}_k)^T$, and let $\lambda_1^k \geqslant \cdots \geqslant \lambda_D^k$ be the eigenvalues of Σ_k. Define $\delta_d^k = \delta_d(\Sigma_k) := \frac{1}{2}(\lambda_d^k - \lambda_{d+1}^k)$. On the other hand, let $\hat{\Sigma}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbf{x}_i - \hat{\mathbf{u}}_k)(\mathbf{x}_i - \hat{\mathbf{u}}_k)^T$. Let $\bar{n} = \min(n_1, \cdots, n_K)$ and $\mathscr{B}(\mathbf{u}_k, M)$ represent a ball with center $\mathbf{u}_k$ and radius M. Eventually, we denote $\|A\|$ as the spectral norm of a matrix A and $\operatorname{tr}(A)$ as its trace and we use I to represent the identity matrix.

2.3 *The Consistency Result*

As mentioned in Section 1, a desirable property for classifiers is *consistency*. Denote h_n to be any classification rule determined from n samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, f^* as the optimal Bayes rule, i.e., $f^* = \arg\min_f R(f)$ and $R^* := R(f^*)$ as its risk. Now we define strong consistency in the following sense.

Definition 1 (Strong Consistency) A classification rule h_n is said to be strongly consistent if

$$R(h_n) \to R^* \text{ a.s.} \quad \text{as } n \to \infty$$

Since the NSS classifer is also based on n samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, from now on, we denote it as $\hat{f}_n$ for it for the rest of the paper. Then we obtain the following theorem for the NSS classifier described in Algorithm 1.

Theorem 1 The NSS classifier $\hat{f}_n$ is strongly consistent, i.e., $R(\hat{f}_n) \to R^*$ a.s. as $n \to \infty$, when the following assumptions hold.

(1) $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ are i.i.d. samples of random variable (X,Y); $X \in \mathbb{R}^D$ and $Y \in \{1, \ldots, K\}$.
(2) $\mathbb{P}(Y = i) = \frac{1}{K}$.
(3) $\mathbb{P}(X|Y = k) = \int_{X \in C_k} g(\mathrm{dist}(X, L_k))dX$, where $g(\cdot)$ decreases exponentially w.r.t. $\mathrm{dist}^2(X, L_k)$.
(4) $\|X\| \leqslant M$.

This theorem reveals that the average prediction error of NSS converges to the optimal prediction error under certain conditions. It is a similar but slightly weaker result in contrast to that for LDA in [57], since the above condition (3) is stronger than that for LDA. Note that both results are about consistency for a class of distributions. On the other hand, the consistency results for KNN, SVM and some boosting methods are for all distributions, and thus are more general [4, 8, 50, 52].

2.4 *The Non-Asymptotic Result*

Theorem 2 With probability at least $1 - 3e^{-s}$, we have

$$0 \leqslant R(\hat{f}_n) - R^* \leqslant C(a, M, M_1, \delta_d)\sqrt{\frac{s + \log(\max(b, 8))}{\bar{n}}}$$

under the following assumptions for some constants C and b. An explicit form of C will be given in the proof and b is defined in the following assumption (6).

(1) $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ are i.i.d. samples of random variable (X,Y); $X \in \mathbb{R}^D$ and $Y \in \{1, \ldots, K\}$.
(2) $\mathbb{P}(Y = i) = \frac{1}{K}$.
(3) $\mathbb{P}(X|Y = k) = \int_{X \in C_k} r(\mathrm{dist}(X, L_k))dX$, where $r(\cdot)$ is a decreasing function.
(4) $r(\cdot)$ is twice differentiable and $r'(\cdot)$, $r''(\cdot) \leqslant M_1$, for $k = 1, \cdots, K$.
(5) $\|X\| \leqslant M$ and if $X \in C_k$, then $\|X - \mathbf{u}_k\| \leqslant a$, for $k = 1, \cdots, K$.
(6) $n_k \geqslant s + \log(\max(b, 8))$ and $8a^2\sqrt{\frac{s+\log(\max(b,8))}{n_k}} \leqslant \frac{\delta_d}{2}$; where $b = b_1 = \cdots = b_K$ and $\delta_d = \delta_d^1 = \cdots = \delta_d^K$, $b_k = 4\frac{\mathrm{tr}(\mathbb{E}Z_k^2)}{\|\mathbb{E}Z_k^2\|}$ and $Z_k = \frac{1}{n_k}[(X - \mathbf{u}_k)(X - \mathbf{u}_k)^T - \Sigma_k]$ for $X \in C_k$. Due to the above assumption (3), both b and δ_d are independent of k.

This theorem provides a non-asymptotic bound for the difference between the average prediction error of NSS and the optimal prediction error. Note that in the assumption (5), the two inequalities are not independent; in fact, one inequality can imply the other. We present both of them in order to obtain a better estimation for the constant C; in real problems a could be much less than M. The details will be found in the proof of Theorem 2. Note that Theorem 2 implies the convergence of $R(\hat{f}_n)$ to R^* (as $n \to \infty$) in probability, which is weaker than that of Theorem 1. This is also consistent with the fact that the the assumption (3) of Theorem 2 is more general than that for Theorem 1.

On the other hand, Theorem 2 shows that the convergence rate for NSS is $\mathscr{O}(n^{-1/2})$. In fact, the best rate for SVM is up to $\mathscr{O}(n^{-1})$ [49, 51] and LDA has convergence rate $\mathscr{O}(n^{-1}\log D))$ [25]. As long as the dimension is not of the exponential order of n, the rate for LDA is better than that for NSS. KNN is proven to be of $\mathscr{O}(n^{-2/D})$ [32]. Therefore NSS has a better rate when D is large. In summary, the convergence rate for NSS is weaker than the best rate obtained by other classification algorithms thus far. It is worth to note that some of the early work [63] give similar rate as $\mathscr{O}(n^{-1/2})$ for SVM.

2.5 *Discussions*

The NSS algorithm is a very simple and basic classification method; it assumes linear structure in data. However, its decision boundary is more complex than linear boundaries. Models like NSS have their limitations. Although we have given examples where real data can be well approximated by multiple subspaces in the introduction, in practice, simple models like NSS often are not satisfied. However, they are important for the following reasons: (1) They are easy to compute and analyze. (2) They often have good interpretations, critical in many applications. (3) They might be the best that can be done when the available training data are limited. (4) They are the foundation from which more complex models can be generalized (see [27] for more discussion). Therefore, it is important to study these simple methods thoroughly, even if in practice they are no longer state-of-the-art.

The NSS method has been modified and extended through different methods: localization, the kernel trick and the hybrid model. The local subspace methods find, for the investigated data sample, their nearest neighbors in each class and attribute by their distances to the subspace spanned by these neighbors [12, 33, 34, 48, 59]. Due to the fact that only an inner product is needed in the NSS algorithm, it can be naturally extended by the kernel trick, where the original data are embedded into a higher dimensional space and subspace structures are learned there [2, 33, 35, 54, 64]; these two techniques are combined in [65]. Another direction is to represent each class by multiple subspaces [33, 34, 53], where [53] also uses a more general metric than the Euclidian distance. All of these extended techniques define nonlinear decision boundaries and the recent works [12, 34, 53] have shown their state-of-the-art performance.

3. Proof of Theorem 1

In this section, we give a complete proof of Theorem 1 following [57].

3.1 *Preliminaries*

We first describe the problem in detail and prepare to prove the theorem. Consider a classification problem, where the goal is to assign an individual instance to one of K classes, given n observations of (X,Y). To do this, the space $\mathbb{R}^D$ is partitioned into subsets $H_1,\dots,H_K$ such that, for $k=1,\dots,K$, the individual instance is classified to be in group k when $X\in H_k$. This procedure generates a discriminant rule as a mapping $f:\mathbb{R}^D\to\{1,\dots,K\}$ that takes the value $f(X)=k$ whenever the individual is assigned to the k^{th} group, and this can be written as $f(X)=\sum_{k=1}^K k\mathbb{1}_{H_k}(X)$, where $\mathbb{1}_{H_k}(X)$ is the indicator function of the subset H_k.

Let Y be the discrete random variable (class index or group label) which represents the true membership of the individual under study. Denote the class prior probabilities $\pi_k=\mathbb{P}[Y=k]>0$, $\sum_{k=1}^K\pi_k=1$ and $k=1,\dots,K$. Furthermore, assume there exist density functions $g_k(X)$ such that $\mathbb{P}[X\in\mathscr{A}|Y=k]=\int_{\mathscr{A}}g_k(X)dX$, $k=1,\dots,K$ for $\mathscr{A}$, a subset of $\mathbb{R}^D$.

Given (X,Y), the rule $f(X)=\sum_{k=1}^K kI_{H_k}(X)$ is in error when $f(X)\neq Y$ and its probability of misclassification is computed as:

$$\begin{aligned} R(f) &= \mathbb{E}_{(X,Y)}\mathbb{1}(f(X)\neq Y)=\mathbb{P}[f(X)\neq Y]=1-\mathbb{P}[f(X)=Y] \\ &= 1-\sum_{k=1}^{K}\mathbb{P}[X\in H_k, Y=k]=1-\sum_{k=1}^{K}\mathbb{P}[Y=k]\mathbb{P}[X\in H_k|Y=k] \\ &= 1-\sum_{k=1}^{K}\pi_k\int_{H_k}g_k(X)dX. \end{aligned} \tag{3.1}$$

The rule $f^*=\sum_{k=1}^{K}k\mathbb{1}_{H_k^*}(X)$ that minimizes (3.1), or the Bayes rule, is given by the partition

$$H_k^*=[X:\ \pi_k g_k(X)=\max_{\mathbb{1}\leqslant j\leqslant K}\pi_j g_j(X)],\ \ k=1,\ldots,K.$$

Then the corresponding optimal error is:

$$R^*=R[f^*(X)]=1-\sum_{k=1}^{K}\pi_k\int_{H_k^*}g_k(X)dX.$$

In general, both π_k and g_k are unknown, so rules used in practice are sample based rules of the form $\hat{f}_n(X)=\sum_{k=1}^{K}kI_{\hat{H}_{k,n}}(X)$, where the subsets $\hat{H}_{k,n}$ depend on the data set $\Omega_n=\{(\mathbf{x}_i,y_i)\}_{i=1}^{n}$ formed by n i.i.d. observations from (X,Y). The appropriate measure of error of a sample rule $\hat{f}_n(X)$ is $R_n=\mathbb{P}[\hat{f}_n(X)\neq Y]$.

3.2 *Proof of Theorem 1*

LEMMA 3.1 Assume $\pi_k=\frac{1}{K}$ and let $\hat{g}_{k,n}(X)$ be an estimate of $g_k(X)$ from Ω_n , for $k=1,\cdots,K$. Let $\hat{f}_n(X)$ be the classifier derived from $\hat{g}_{k,n}(X)$, i.e., $\hat{f}_n(X)=\arg\max_k\hat{g}_{k,n}(X)$. Then

$$0\leqslant R_n-R^*\leqslant\frac{1}{K}\sum_{k=1}^{K}\int|g_k(X)-\hat{g}_{k,n}(X)|dX.$$

This lemma gives a useful bound for $\leqslant R_n-R^*$. A similar result of Lemma 3.1 can be found in the Theorem 1 in [20] (p. 254). We provide our proof of Lemma 3.1 in Appendix A.

Proof of Theorem 1. Due to condition (2), we have

$$H_k^*=[X:g_k(X)=\max_{1\leqslant j\leqslant K}g_j(X)].$$

On the other hand, based on the assumption (3), the density functions can be written as

$$\begin{aligned} &g_k(t)=C_1(d)\exp(-\alpha t),\\ &t=(X-\mathbf{u}_k)^T(I-\mathbf{P}_k)(X-\mathbf{u}_k) \end{aligned}$$

for some $\alpha>0$ and constant $C(d)$ and $\mathbf{P}_k=\mathbf{B}_k\mathbf{B}_k^T$ with $\mathbf{B}_k$ being the orthonormal basis for L_k.

Then the classifier generated by the Algorithm 1 can be written as:

$$\hat{f}_n(X)=\sum_{k=1}^{K}kI_{\hat{H}_{k,n}}$$

with the following notation:

$$\begin{aligned}
&\hat{\mathbf{P}}_k = \hat{\mathbf{B}}_k \hat{\mathbf{B}}_k^T \\
&\hat{g}_{k,n}(x) = C_1(d) \exp\left(-\alpha(X-\hat{\mathbf{u}}_k)^T(I-\hat{\mathbf{P}}_k)(X-\hat{\mathbf{u}}_k)\right) \\
&\hat{H}_{k,n} = [X : \hat{g}_{k,n}(X) = \max_{1\leqslant j\leqslant K} \hat{g}_{j,n}(X)]
\end{aligned}$$

Thus the NSS classifier can be considered as a plug-in version of the Bayes rule. By Lemma 3.1, the difference $R_n - R^*$ can be bounded in the form

$$0 \leqslant R_n - R^* \leqslant \frac{1}{K}\sum_{k=1}^{K}\int_{\mathbb{R}^D} |g_k(X) - \hat{g}_{k,n}(X)|dX$$

For each fixed $1 \leqslant k \leqslant K$, we have

$$0 \leqslant \int_{\mathbb{R}^D} |g_k(X) - \hat{g}_{k,n}(X)|dX \leqslant \int_{\mathbb{R}^D} g_k(X) + \hat{g}_{k,n}(X)dX < \infty$$

Therefore, it suffices to show that $\hat{g}_{k,n} \to g_k$ *a.s* and due to the continuity of $g(\cdot)$, to show $\hat{\mathbf{u}}_k \to \mathbf{u}_k$ and $\hat{\mathbf{P}}_k \to \mathbf{P}_k$ *a.s.* The fact that $\hat{\mathbf{u}}_k$ and $\hat{\mathbf{P}}_k$ are the maximum-likelihood estimations (MLE) of $\mathbf{u}_k$ and $\mathbf{P}_k$ completes the proof. □

4. Proof of Theorem 2

In this section, we give a complete proof of Theorem 2.

4.1 *Preliminary Results*

We first present several lemmas that will lead to Theorem 2. For the following lemmas, we make assumptions (1) and (5) of Theorem 2.

LEMMA 4.1 ([36], Lemma 11) For all s such that $s + \log(8) \leqslant n_k$, with probability $\geqslant 1 - e^{-s}$, we have

$$\|\mathbf{u}_k - \hat{\mathbf{u}}_k\| \leqslant 2a\sqrt{\frac{s+\log(8)}{n_k}}.$$

LEMMA 4.2 For all s such that $s + \log(\max(b,8)) \leqslant n_k$ and $8a^2\sqrt{\frac{s+\log(\max(b,8))}{n_k}} \leqslant \frac{\delta_d}{2}$, with probability at least $\geqslant 1 - 2e^{-s}$, we have

$$\|\mathbf{P}_k - \hat{\mathbf{P}}_k\| \leqslant \frac{8a^2}{\delta_d}\sqrt{\frac{s+\log(\max(b,8))}{n_k}}.$$

A very similar version of Lemma 4.2 can be found in [36]. Our version treats the constants slightly different. For completeness, we will give a proof of Lemma 4.2 in Appendix B.

LEMMA 4.3 Assume that X belongs to the k^{th} class. For all s such that $s + \log(\max(b,8)) \leqslant n_k$ and $8a^2\sqrt{\frac{s+\log(\max(b,8))}{n_k}} \leqslant \frac{\delta_d}{2}$, with probability at least $\geqslant 1 - 3e^{-s}$, we have

$$|\operatorname{dist}(X, L_k) - \operatorname{dist}(X, \hat{L}_k)| \leqslant (2a + \frac{16a^2M}{\delta_d})\sqrt{\frac{s+\log(\max(b,8))}{n_k}}.$$

where $\mathrm{dist}(X,\hat{L}_k) := \|(I-\hat{\mathbf{P}}_k)(X-\hat{\mathbf{u}}_k)\|$.

Proof.

$$\begin{aligned}
\mathrm{dist}(X,L_k)-\mathrm{dist}(X,\hat{L}_k) &= \|(I-\mathbf{P}_k)(X-\mathbf{u}_k)\|-\|(I-\hat{\mathbf{P}}_k)(X-\hat{\mathbf{u}}_k)\| \\
&= \|(I-\hat{\mathbf{P}}_k+\hat{\mathbf{P}}_k-\mathbf{P}_k)(X-\hat{\mathbf{u}}_k+\hat{\mathbf{u}}_k-\mathbf{u}_k)\|-\|(I-\hat{\mathbf{P}}_k)(X-\hat{\mathbf{u}}_k)\| \\
&\leqslant \|(I-\hat{\mathbf{P}}_k)(\hat{\mathbf{u}}_k-\mathbf{u}_k)\|+\|(\hat{\mathbf{P}}_k-\mathbf{P}_k)(X-\hat{\mathbf{u}}_k)\| \\
&\leqslant \|\hat{\mathbf{u}}_k-\mathbf{u}_k\|+2M\|\hat{\mathbf{P}}_k-\mathbf{P}_k\|
\end{aligned}$$

Applying Lemma 4.1 and 4.2 completes the proof. □

4.2 *Proof of Theorem 2*

Now we prove Theorem 2 in this section.

Proof of Theorem 2. Let $\hat{r}_{k,n} := r(\mathrm{dist}(X,\hat{L}_k))$. Then NSS assigns a sample $\mathbf{x}$ to the class $\arg\min_k \mathrm{dist}(\mathbf{x},\hat{L}_k))$, i.e., to the class $\arg\max_k \hat{r}_{k,n}$. Therefore, by Lemma 3.1, we know

$$0 \leqslant R_n - R^* \leqslant \frac{1}{K}\sum_{k=1}^{K}\int |r_k(\mathrm{dist}(X,L_k))-r_k(\mathrm{dist}(X,\hat{L}_k))|dX.$$

Now we need to analyze $|r_k(\mathrm{dist}(X,L_k))-r_k(\mathrm{dist}(X,\hat{L}_k))|$. Let $t_k = \mathrm{dist}(X,L_k)$ and $\hat{t}_k = \mathrm{dist}(X,\hat{L}_k)$. Then the Taylor's theorem gives

$$r_k(t_k) = r_k(\hat{t}_k)+r_k'(\hat{t}_k)(t_k-\hat{t}_k)+\frac{r_k''(a_0)}{2}(t_k-\hat{t}_k)^2$$

for some number a_0 between t_k and $\hat{t}_k$. Thus

$$|r_k(t_k)-r_k(\hat{t}_k)| \leqslant M_1|t_k-\hat{t}_k|+\frac{M_1}{2}|t_k-\hat{t}_k|^2$$

By Lemma 4.3 and the fact that $s+\log(\max(b,8)) \leqslant n_k$, we have

$$|r_k(t_k)-r_k(\hat{t}_k)| \leqslant 2M_1[(a+\frac{8a^2M}{\delta_d})+(a+\frac{8a^2M}{\delta_d})^2]\sqrt{\frac{s+\log(\max(b,8))}{n_k}}$$

Therefore,

$$0 \leqslant R_n - R^* \leqslant 2KVM_1[(a+\frac{8a^2M}{\delta_d})+(a+\frac{8a^2M}{\delta_d})^2]\sqrt{\frac{s+\log(\max(b,8))}{\bar{n}}}$$

where V is the volume of the domain of X and is $\leqslant M^D$. Putting $C(a,M,M_1,\delta_d) = 2KVM_1[(a+\frac{8a^2M}{\delta_d})+(a+\frac{8a^2M}{\delta_d})^2]$ completes the proof. □

5. Experiments

In this section, we present some experiments that are related to our theoretical results. Our experiments consist of two parts. In the first part, the simulated data generally follow the assumptions of our theorems (or at least do so with high probability), while the data in the other part do not.

5.1 *Simulations under the Assumptions of Theorem 1 and 2*

5.1.1 *Convergence Rate.*

DATA. Our data are simulated as follows. Let $D = 3$, $K = 2$. We generate our samples from two Gaussians with means $\mu_1 = (1,1,1)^T$, $\mu_2 = (-1,-1,-1)^T$ and variances

$$\Sigma_1 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \Sigma_2 = U \begin{pmatrix} 2 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} U^T,$$

where $U^T U = UU^T = I$. For the training set, we generate n_1 and n_2 points for the two classes respectively. So the total number of training samples is $n = n_1 + n_2$. For the testing set, we do so for N_1 and N_2 points. We apply the NSS algorithm with $d = 1$. The testing set is used to compute the risk $R(\hat{f}_n)$ and R^*. We let $\lambda = 0.5$ and $n_1 = n_2$ vary from 5 to 500.

RESULT. Since our theoretical result gives a non-asymptotic bound with high probability. We repeat the process of generating the training data and learning the NSS classifier 100 times and obtain 100 measurements for $R(\hat{f}_n)$. Then we plot the minimum, 25%, 50% and 75% percentiles of $R(\hat{f}_n) - R^*$ against $\sqrt{n}$ in the left of Figure 1. The repetition procedure is run for all the experiments in this paper; whenever we plot $R(\hat{f}_n) - R^*$, we always plot these percentiles. From the left figure of Figure 1, we can observe that $R(\hat{f}_n) - R^*$ actually decreases faster than the order of $n^{-1/2}$.

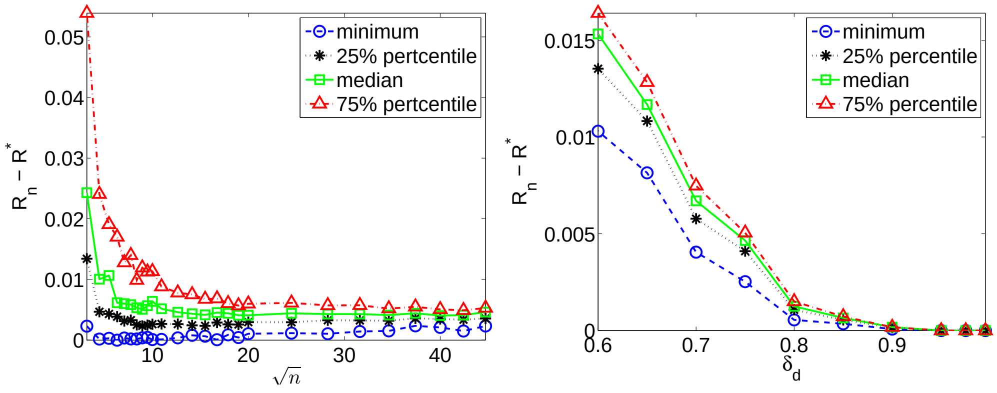

FIG. 1: Change of $R(\hat{f}_n) - R^*$ w.r.t. $\sqrt{n}$ on the left and w.r.t. δ_d on the right

5.1.2 *Effect of the Eigengap δ_d.*

DATA. We follow Section 5.1.1 to generate our data. However, we fix $n = 1000$ and vary λ from 0.01 to 0.8.

RESULT. We compute the eigengap $\delta_d = (2-\lambda)/2$ and plot $R(\hat{f}_n) - R^*$ against δ_d in the right of Figure 1. It can be seen that $R(\hat{f}_n) - R^*$ decreases as δ_d increases; the rate of change is close to but slower than the linear rate.

5.1.3 *Effect of the Ambient Dimension D.*

DATA. We follow Section 5.1.1 to generate our data. However, we fix $n = 1000$ and vary D from 3 to 20. To ensure the same separation between classes we let $\mu_1 = (1,\cdots,1)^T$, $\mu_2 = (-1,-1,-1,1,\cdots,1)^T$. Our variances follow the same form as in Section 5.1.1, i.e., the first eigenvalue are 2 and the rest are $\lambda = 0.5$.

RESULT. The quantity $R(\hat{f}_n) - R^*$ versus D is displayed in the left of Figure 2. We see that NSS favors the high dimensions while both the eigengap δ_d and the separation between classes are fixed.

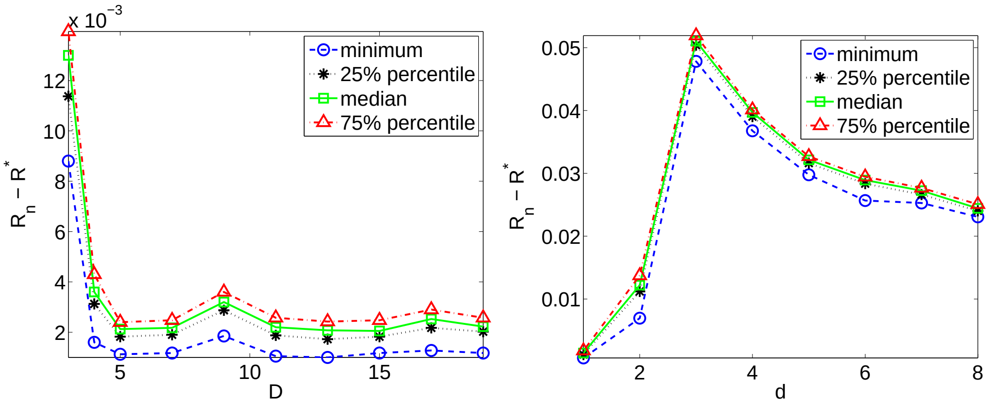

FIG. 2: Change of $R(\hat{f}_n) - R^*$ w.r.t. D on the left and w.r.t. d on the right

5.1.4 *Effect of the Intrinsic Dimension d.*

DATA. We follow Section 5.1.1 to generate our data. However, we fix $n = 1000$, $D = 20$ and vary d from 1 to 8. To ensure the same separation between classes we let $\mu_1 = (1,\cdots,1)^T$, $\mu_2 = (-1,-1,-1,1,\cdots,1)^T$. Our variances follow the same form as in Section 5.1.1, i.e., the first d eigenvalues is 2 and the rest $D-d$ eigenvalues are $\lambda = 0.5$.

RESULT. The quantity $R(\hat{f}_n) - R^*$ versus d is displayed in the right of Figure 2. It can be seen that $R(\hat{f}_n) - R^*$ increases w.r.t. d at first and then starts to drop from $d = 4$. Our understanding is that learning the projector of a subspace with higher intrinsic dimension requires more training samples to obtain the same accuracy. However, on the other hand, while the intrinsic dimension increases, the noise decreases in our setting. This tradeoff explains the turning point in the right figure of Figure 2

5.2 *Simulations in More General Conditions*

5.2.1 *Different* $\mathbb{P}(X|Y=k)$.

DATA. We follow Section 5.1.1 to generate our data. However, our covariance matrices are:

$$\Sigma_1 = \begin{pmatrix} 1.5 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \Sigma_2 = U \begin{pmatrix} 2.5 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} U^T,$$

RESULT. The quantity $R(\hat{f}_n) - R^*$ versus $\sqrt{n}$ is displayed in the left of Figure 3. We can observe that $R(\hat{f}_n)$ still converges to R^* at a rate faster than $n^{-1/2}$.

5.2.2 *Unequal* $\mathbb{P}(Y=k)$.

DATA. We follow Section 5.1.1 to generate our data. However, in this experiment, we let $n_2 = 3n_1$ and let n_1 vary from 5 to 500.

RESULT. We plot $R(\hat{f}_n) - R^*$ against $\sqrt{n}$ in the right of Figure 3. It can be seen that $R(\hat{f}_n)$ still converges to R^* at a rate faster than $n^{-1/2}$.

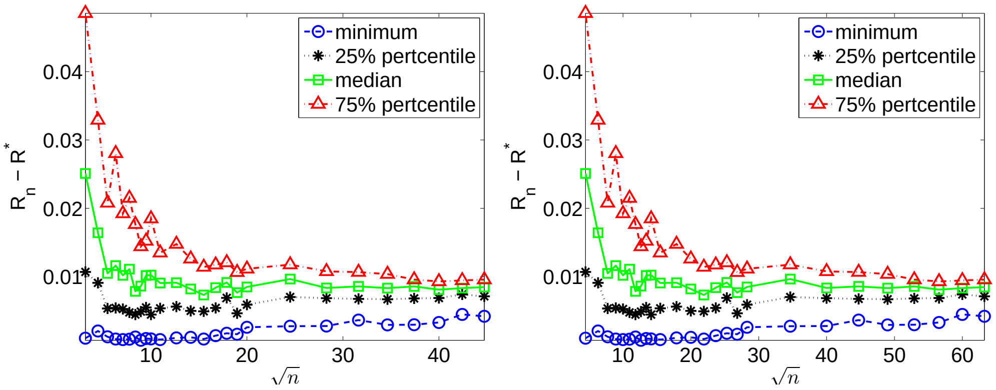

FIG. 3: Change of $R(\hat{f}_n) - R^*$ w.r.t. $\sqrt{n}$. On the left, $\mathbb{P}(X|Y=k)$ are different ; on the right, $\mathbb{P}(Y=k)$ are not equal

5.3 *Discussions*

To summarize, we observe faster convergence rate in the above example than that in Theorem 2.The effects of the ambient dimension D and the intrinsic dimension d are very interesting and they could be a guide for practice. Moreover, the last two examples demonstrate that the NSS classifier still converges to the Bayes rule even when some of the assumptions in our theorems are not satisfied. Last but not least, the separation between the classes should affect the convergence and it is not in our current bound yet.

These experiments practically verify our theoretical results and motivate us to improve our theoretical results further in the future.

6. Conclusion

In this paper, we reviewed a simple classification algorithm (NSS) based on the model of multiple subspaces. We proved its strong consistency under certain conditions, which means that under these conditions, the prediction error of NSS on average converges strongly to that of the optimal classifier. Other than this, we provided a non-asymptotic bound for the difference between the NSS risk and the Bayes risk (we will call this different "error"). This result tells how many data points are required for obtaining an error that is less than a certain small number with high probability. Our simulations also provide many interesting observations and practically verify the theoretical results.

By studying the consistency property of NSS, we are inspired to further explore subspace-based classification methods along the following directions in the future. First, NSS finds a good estimation for the underlying subspace models by minimizing the sum of squares of fitting errors. However, for the purpose of classification, it is more helpful to obtain models which can "separate" or "discriminate" classes. Therefore, in order to improve the classification performance, some separation measure can be taken into account. In fact, an advanced supervised learning method based on multiple subspaces has been proposed [53]. It would be fruitful to analyze this method or other variants theoretically.

Moreover, a general way to find a good classifier is to minimize an empirical risk function, which is typically defined as $R_{\text{emp}}(f) = \sum_{i=1}^{n} \mathbb{1}(f(\mathbf{x}_i) \neq y_i)$. This idea can be combined with the multiple subspaces model. Similar approaches to that in [56] can be applied to analyze its consistency.

Finally, our experiments suggest that it is promising to obtain convergence rate that is faster than $\mathscr{O}(n^{-1/2})$. It is worth to explore this direction further for an improved rate.

Acknowledgements

The author thanks Dr. Gilad Lerman for motivating her to work on the problem and for valuable discussions. We are also grateful to Dr. Mauro Maggioni for his precious guidance on the non-asymptotic results.

A. Proof of Lemma 3.1

Proof. Since $\pi_k = \frac{1}{K}$, we have $H_k^* = [X : \ g_k(X) = \max_{1 \leqslant j \leqslant K} g_j(X)$. Thus,

$$\begin{aligned} R^* &= 1 - \frac{1}{K} \sum_{k=1}^{K} \int_{H_k^*} g_k(X) dX = 1 - \frac{1}{K} \int \max_k g_k(X) dX \\ &\leqslant 1 - \frac{1}{K} \int g_{\hat{f}_n}(X) dX = R_n \end{aligned}$$

On the other hand,

$$\begin{aligned} R_n - R^* &= \frac{1}{K}\int(\max_k g_k(X) - g_{\hat{f}_n}(X))dX \\ &= \frac{1}{K}\int(\max_k g_k(X) - \hat{g}_{\hat{f}_n,n}(X))dX + \frac{1}{K}\int(\hat{g}_{\hat{f}_n,n}(X) - g_{\hat{f}_n}(X))dX \\ &= \frac{1}{K}\int(\max_k g_k(X) - \max_k \hat{g}_{k,n}(X))dX + \frac{1}{K}\int(\hat{g}_{\hat{f}_n,n}(X) - g_{\hat{f}_n}(X))dX \\ &\leqslant \frac{1}{K}\sum_{k=1}^{K}\int|g_k(X) - \hat{g}_{k,n}(X)|dX. \end{aligned}$$

□

B. Proof of Lemma 4.2

Theorem 3 ([37], Theorem 2.1) Let $Z_1,\cdots,Z_n \in \mathbb{R}^{D\times D}$ be a sequence of independent symmetric random matrices such that $\mathbb{E}Z_i = 0$ and $\|Z_i\| \leqslant U$ a.s., $1 \leqslant i \leqslant n$. Let

$$\sigma^2 := \left\|\sum_i^n \mathbb{E}Z_i^2\right\|$$

Then for any $s \geqslant 1$,

$$\left\|\sum_i^n \mathbb{E}Z_i\right\| \leqslant 2\max\left(\sigma\sqrt{t+\log(\bar{B})}, U(t+\log(\bar{B}))\right)$$

with probability at least $1-e^{-s}$, where $\bar{B} := 4\,\mathrm{tr}\left(\sum_{i=1}^n \mathbb{E}Z_i^2\right)/\sigma^2$.

Theorem 4 ([19], [66]) If $\|\hat{\Sigma}_k - \Sigma_k\| \leqslant \delta_d^k/2$, then

$$\left\|\mathbf{P}_k - \hat{\mathbf{P}}_k\right\| \leqslant \frac{\|\hat{\Sigma}_k - \Sigma_k\|}{\delta_d^k}.$$

Now we use Theorem 3 and 4 to prove Lemma 4.2 following [36].

Proof of Lemma 4.2. First, we show that with probability at least $1-2e^{-s}$,

$$\|\hat{\Sigma}_k - \Sigma_k\| \leqslant 8a^2\sqrt{\frac{s+\log(\max(b_k,8))}{n_k}} \tag{A.1}$$

Since

$$\begin{aligned} \|\hat{\Sigma}_k - \Sigma_k\| &= \|\frac{1}{n_k}\sum_{\mathbf{x}_i\in C_k}(\mathbf{x}_i-\mathbf{u}_k)(\mathbf{x}_i-\mathbf{u}_k)^T - (\mathbf{u}_k-\hat{\mathbf{u}}_k)(\mathbf{u}_k-\hat{\mathbf{u}}_k)^T - \Sigma_k\| \\ &= \|\frac{1}{n_k}\sum_{\mathbf{x}_i\in C_k}(\mathbf{x}_i-\mathbf{u}_k)(\mathbf{x}_i-\mathbf{u}_k)^T - \Sigma_k\| + \|(\mathbf{u}_k-\hat{\mathbf{u}}_k)(\mathbf{u}_k-\hat{\mathbf{u}}_k)^T\| \end{aligned}$$

Let $Z_{k_i} = \frac{1}{n_k}\left[(\mathbf{x}_{k_i}-\mathbf{u}_k)(\mathbf{x}_{k_i}-\mathbf{u}_k)^T - \Sigma_k\right]$, where $\mathbf{x}_{k_i} \in C_k$. For simplicity, we write Z_i for Z_{k_i} and $\mathbf{x}_i$ for $\mathbf{x}_{k_i}$ for the rest of the proof. Then, we have

$$\begin{aligned} \mathbb{E}Z_i &= 0, \quad \|Z_i\| \leqslant \frac{1}{n_k}(\|(\mathbf{x}_i-\mathbf{u}_k)(\mathbf{x}_i-\mathbf{u}_k)^T\| + \|\Sigma_k\|) \leqslant \frac{1}{n_k}(a^2+\|\Sigma_k\|) \\ \|\Sigma_k\| &= \|\mathbb{E}_{X\in C_k}(X-\mathbf{u}_k)(X-\mathbf{u}_k)^T\| \leqslant \mathbb{E}_{X\in C_k}(\|(X-\mathbf{u}_k)(X-\mathbf{u}_k)^T\|) \leqslant a^2 \end{aligned}$$

Thus we have

$$\|Z_i\| \leqslant \frac{2a^2}{n_k}$$

Moreover,

$$\begin{aligned} \sigma^2 &:= \|\sum_{i=1}^{n_k}\mathbb{E}Z_i^2\| = \|\sum_{i=1}^{n_k}\mathbb{E}\frac{1}{n_k^2}\left[(\mathbf{x}_i-\mathbf{u}_k)(\mathbf{x}_i-\mathbf{u}_k)^T - \mathbb{E}(\mathbf{x}_i-\mathbf{u}_k)(\mathbf{x}_i-\mathbf{u}_k)^T\right]^2\| \\ &= \frac{1}{n_k}\|\mathbb{E}\left[(\mathbf{x}_1-\mathbf{u}_k)(\mathbf{x}_1-\mathbf{u}_k)^T\right]^2 - \Sigma_k^2\| \leqslant \frac{1}{n_k}\left(\|\mathbb{E}[(\mathbf{x}_1-\mathbf{u}_k)(\mathbf{x}_1-\mathbf{u}_k)^T\right]^2\| + \|\Sigma_k^2\|\right) \\ &\leqslant \frac{2a^4}{m} \end{aligned}$$

Applying Theorem 3 gives

$$\begin{aligned} \|\frac{1}{n_k}\sum_{\mathbf{x}_i\in C_k}(\mathbf{x}_i-\mathbf{u}_k)(\mathbf{x}_i-\mathbf{u}_k)^T - \Sigma_k\| &= \|\sum_i^{n_k} Z_i\| \leqslant 2\max\left(\sqrt{\frac{2}{n_k}}a^2\sqrt{s+\log(b_k)}, \frac{2a^2}{n_k}\big(s+\log(b_k)\big)\right) \\ &= 2a^2\sqrt{\frac{2\big(s+\log(b_k)\big)}{n_k}}\max(1,\sqrt{\frac{2\big(s+\log(b_k)\big)}{n_k}}) \\ &\leqslant 4a^2\sqrt{\frac{s+\log b_k}{n_k}} \end{aligned}$$

where $b_k = 4\,\mathrm{tr}\left(\sum_{i=1}^{n_k}\mathbb{E}Z_i^2\right)/\sigma^2 = 4\frac{tr(\mathbb{E}Z_1^2)}{\|\mathbb{E}Z_1^2\|}$. Combining this result and Lemma 4.1 implies (A.1). The fact that $8a^2\sqrt{\frac{s+\log(\max(b_k,8))}{n_k}} \leqslant \frac{\delta_d^k}{2}$ and Theorem 4 complete the proof.

□

REFERENCES

[1] BACH, F. & AUDIBERT, J.-Y. (2008) Supervised learning for computer vision: Theory and algorithms. .

[2] BALACHANDER, T. & KOTHARI, R. (1999) Kernel based subspace pattern classification. in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

[3] BARTLETT, P. L., JORDAN, M. I. & MCAULIFFE, J. D. (2006) Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, **101**(473), 138–156.

[4] BARTLETT, P. L. & TRASKIN, M. (2007) AdaBoost is Consistent.. *Journal of Machine Learning Research*, **8**, 2347–2368.

[5] BASRI, R. & JACOBS, D. (2003) Lambertian Reflectance and Linear Subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(2), 218–233.

[6] Bishop, C. M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[7] Blanchard, G., Bousquet, O. & Massart, P. (2008) Statistical performance of support vector machines. *The Annals of Statistics*, **36**(2), 489–531.

[8] Blanchard, G., Lugosi, G. & Vayatis, N. (2003) On the Rate of Convergence of Regularized Boosting Classifiers. *J. Mach. Learn. Res.*, **4**, 861–894.

[9] Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992) A Training Algorithm for Optimal Margin Classifiers. in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, New York, NY, USA. ACM.

[10] Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

[11] Cappelli, R., Maio, D. & Maltoni, D. (2001) Multispace KL for Pattern Representation and Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(9), 977–996.

[12] Cevikalp, H., Larlus, D., Neamtu, M., Triggs, B. & Jurie, F. (2010) Manifold Based Local Classifiers: Linear and Nonlinear Approaches.. *Signal Processing Systems*, **61**(1), 61–73.

[13] Chaudhuri, K. & Dasgupta, S. (2014) Rates of Convergence for Nearest Neighbor Classification. in *Advances in Neural Information Processing Systems (NIPS) 27*, pp. 3437–3445.

[14] Chen, D., Wu, Q., Ying, Y. & Zhou, D. (2004) Support Vector Machine Soft Margin Classifiers: Error Analysis.. *Journal of Machine Learning Research*, **5**, 1143–1175.

[15] Cortes, C. & Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**(3), 273–297.

[16] Costeira, J. & Kanade, T. (1998) A Multibody Factorization Method for Independently Moving Objects. *International Journal of Computer Vision*, **29**(3), 159–179.

[17] Cover, T. & Hart, P. (2006) Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, **13**(1), 21–27.

[18] Cover, T. M. (1968) Rates of Converence for Nearest Neighbor Procedures. in *Proc. 1st Ann. Hawaii Conf. on Systems Theory*, pp. 413–415.

[19] Davis, C. & Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation, III. *SIAM J. Numer. Anal.*, **7**.

[20] Devroye, L. & Györfi, L. (1985) *Nonparametric density estimation: the L_1 view*. John Wiley, New York.

[21] Duda, R. O. & Hart, P. E. (1973) *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, NY.

[22] Epstein, R., Hallinan, P. & Yuille, A. (1995) 5 ± 2 eigenimages suffice: an empirical investigation of low-dimensional lighting models. in *Physics-Based Modeling in Computer Vision, 1995., Proceedings of the Workshop on*, p. 108.

[23] Fisher, R. A. (1936) The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**(7), 179–188.

[24] Fix, E. & Jr (1951) Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties. Discussion Paper Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, Texas.

[25] Girard, R. (2011) Fast rate of convergence in high-dimensional linear discriminant analysis. *Journal of Nonparametric Statistics*, **23**.

[26] Györfi, L. (1981) The Rate of Convergence of k_n-NN Regression Estimates and Classification Rules. *IEEE Transactions on Information Theory*, **27**(3).

[27] Hastie, T., Tibshirani, R. & Friedman, J. H. (2001) *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag.

[28] Ho, J., Yang, M., Lim, J., Lee, K. & Kriegman, D. (2003) Clustering appearances of objects under varying illumination conditions. in *Proceedings of International Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 11–18.

[29] Kirby, M. & Sirovich, L. (1990) Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, **12**(1), 103–108.

[30] KOHONEN, T., NÉMETH, G., BRY, K.-J., JALANKO, M. & RIITTINEN, H. (1979) Spectral classification of phonemes by learning subspaces. in *ICASSP*, pp. 97–100.

[31] KOHONEN, T., RIITTINEN, H., JALANKO, M., REUHKALA, E. & HALTSONEN, S. (1980) A thousand-word recognition system based on the learning subspace method and redundant hash addressing. in *Proceedings of the 5th Intertional Conference on Pattern Recognition*, pp. 158–165.

[32] KULKARNI, S. R. & POSNER, S. E. (1995) Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, **41**(4).

[33] LAAKSONEN, J. (1997) Subspace Classifiers in Recognition of Handwritten Digits. Ph.D. thesis, Helsinki University of Technology.

[34] LIU, Y., GE, S. S., LI, C. & YOU, Z. (2011) k-NS: A Classifier by the Distance to the Nearest Subspace.. *IEEE Transactions on Neural Networks*, **22**(8), 1256–1268.

[35] MAEDA, E. & MURASE, H. (2002) Kernel-Based Nonlinear Subspace Method for Pattern Recognition.. *Systems and Computers in Japan*, **33**(1), 38–52.

[36] MAGGIONI, M., MINSKER, S. & STRAWN, N. (2014) Multiscale Dictionary Learning: Non-Asymptotic Bounds and Robustness. arXiv:1401.5833v2.

[37] MINSKER, S. (2013) On Some Extensions of Bernstein's Inequality for Self-adjoint Operators. arXiv:1112.5448v2.

[38] MINSKY, M. L. & PAPERT, S. (1988) *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge Mass., expanded ed. edn.

[39] MITCHELL, T. M. (1997) *Machine Learning*. McGraw-Hill, New York.

[40] OJA, E. (1983) *Subspace methods of pattern recognition*, Electronic & electrical engineering research studies. Research Studies Press.

[41] OKUN, O. (2004) Protein Fold Recognition with K-Local Hyperplane Distance Nearest Neighbor Algorithm. in *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics*.

[42] PORIKLI, F. & CHI, Y. (2012) Connecting the dots in multi-class classification: From nearest subspace to collaborative representation. *IEEE Conference on Computer Vision and Pattern Recognition*, **0**, 3602–3609.

[43] QUINLAN, J. R. (1979) Discovering Rules by Induction from Large Collections of Examples. in *Expert Systems in the Micro-Electronic Age*, ed. by D. Michie, pp. 168–201. Edinburgh University Press, Edinburgh.

[44] ——— (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[45] RAO, C. (1952) *Advanced statistical methods in biometric research*, Wiley publications in statistics. Wiley.

[46] ROSENBLATT, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, **65**(6), 386–408.

[47] SCHÖLKOPF, B., BURGES, C. & VAPNIK, V. (1995) Extracting support data for a given task. in *First International Conference on Knowledge Discovery and Data Mining*, Menlo Park. AAAI Press.

[48] SKARBEK, W., GHUWAR, M. & IGNASIAK, K. (1997) Local Subspace Method for Pattern Recognition.. in *CAIP*, ed. by G. Sommer, K. Daniilidis, & J. Pauli, vol. 1296 of *Lecture Notes in Computer Science*, pp. 527–534. Springer.

[49] SRIDHARAN, K., SHALEV-SHWARTZ, S. & SREBRO, N. (2009) Fast Rates for Regularized Objectives. in *Advances in Neural Information Processing Systems (NIPS) 21*, pp. 1545–1552.

[50] STEINWART, I. (2002) Support Vector Machines are Universally Consistent. *J. Complexity*, **18**(3), 768–791.

[51] STEINWART, I. & SCOVEL, C. (2007) Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, **35**(2), 575–607.

[52] STONE, C. J. (1977) Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595–620.

[53] SZLAM, A. & SAPIRO, G. (2009) Discriminative *k*-Metrics. in *Proceedings of the 26th International Conference on Machine Learning*, ed. by L. Bottou, & M. Littman, pp. 1009–1016, Montreal. Omnipress.

[54] TSUDA, K. (1999) Subspace classifier in reproducing kernel hilbert space. in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

[55] TURK, M. & PENTLAND, A. (1991) Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, **3**(1), 71–86.

[56] VAPNIK, V. N. (1999) An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, **10**(5), 988–999.

[57] VELILLA, S. & HERNÁNDEZ, A. (2005) On the Consistency Properties of Linear and Quadratic Discriminant Analyses. *Journal of Multivariate Analysis*, **96**(2), 219–236.

[58] VERT, R. & VERT, J. (2006) Consistency and Convergence Rates of One-Class SVMs and Related Algorithms. *J. Mach. Learn. Res.*, **7**, 817–854.

[59] VINCENT, P. & BENGIO, Y. (2001) K-Local Hyperplane and Convex Distance Nearest Neighbor Algorithms.. in *Advances in Neural Information Processing Systems (NIPS)*, ed. by T. G. Dietterich, S. Becker, & Z. Ghahramani, pp. 985–992. MIT Press.

[60] WATANABE, S., LAMBERT, P. F., KULIKOWSKI, C. A., BUXTON, J. & WALKER, R. (1967) Evaluation and selection of variables in pattern recognition. in *Computer and information sciences*, ed. by J. Tou, vol. 2, pp. 91–122. Academic Press, New York.

[61] WERBOS, P. J. (1974) Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Ph.D. thesis, Harvard University.

[62] WU, Q. & ZHOU, D. (2005) SVM soft margin classifiers: linear programming versus quadratic programming. *Neural Comp*, **17**, 1160–1187.

[63] ZHANG, T. (2003) Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, 56–134.

[64] ZHAO, W. (1999) Subspace methods in object/face recognition. in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

[65] ZOU, D. (2000) Local Subspace Classifier in Reproducing Kernel Hilbert Space.. in *ICMI*, ed. by T. Tan, Y. Shi, & W. Gao, vol. 1948 of *Lecture Notes in Computer Science*, pp. 434–441. Springer.

[66] ZWALD, L. & BLANCHARD, G. (2005) On the Convergence of Eigenspaces in Kernel Principal Component Analysis. in *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pp. 1649–1656.