Nearest Subspace Classification with Missing Data

Yuejie Chi
Electrical and Computer Engineering
The Ohio State University

***Abstract*—We consider the problem of multi-class classification when there are missing entries in both the training samples and the test samples. A modified version of the nearest subspace classifier is proposed and analyzed to handle missing data. We show the performance of the nearest subspace classifier is close to its counterpart when no missing data are present as long as the probability of observing each entry in the training set is $\delta \gtrsim \mathcal{O}((\log M/n_i)^{1/2})$, where M is the sample dimension and $n_i \gtrsim \mathcal{O}(\log M)$ is the training size of the ith class. Finally, numerical results are provided for digit recognition when only a subset of the pixels are observed.**

***Index Terms*—nearest subspace, missing data, multi-class classification**

I. Introduction

Multi-class classification is one of the most important research topics in machine learning, with applications ranging from computer vision, microarray analysis, to signal processing. Conventional algorithms such as the nearest subspace classifier (NSC) [1] and the recently proposed sparse representation based classifier (SRC) [2] have proved to be successful in many cases by assuming samples in the same class lie in a low-dimensional subspace. Compressive Sensing (CS) [3], [4] makes it possible to handle missing entries in the test samples [2], [5], but in many cases it is still required that the training samples are fully observed in order to faithfully extract low-dimensional features.

However, the cost of obtaining complete data may become prohibitively expensive, if not impossible, due to the increasing dimensionality of the datasets of interest in the so-called data deluge. For example, in computational biology, the assessment of a protein or a DNA sequence requires experiments that take a lot of time and resources. Another example is that the privacy settings of users in social networks make certain information unavailable. Therefore there is a demanding need to develop multi-class classification algorithms that don't require complete information of both the training samples and the test samples. Recent advances in low-dimensional manifold modeling of high-dimensional data provide premise for estimating and tracking the structure of the dataset from incomplete observations, such as [6]–[10]. Besides the attempts in estimating the data structure, it is shown that a matched subspace detector can succeed with high probability even with a small number of partial observations under some mild conditions [11].

In this paper, we propose a modified version of the NSC, dubbed the robust nearest subspace classifier (RNSC), to handle missing entries in *both* test and training samples. The proposed algorithm first infer the principal subspace of each class from the partially observed training set, then projects the partially observed testing sample onto the principal subspace of each class, and identifies the class with the minimal residual. When data is fully observed, the proposed RNSC algorithm is the same as the original NSC algorithm. When the probability of observing each entry in the training set is assumed known as δ, we show that the performance of the RNSC is close to the original NSC without missing data as long as $\delta \gtrsim \mathcal{O}((\log M/n_i)^{1/2})$ under mild conditions, where M is the sample dimension and $n_i \gtrsim \mathcal{O}(\log M)$ is the training size of each class. When δ is unknown, we simply assume $\delta = 1$, and this is equivalent to the original NSC by filling in all missing entries as zero. It is empirically shown in numerical examples there is only a very small degeneration in performance.

The rest of the paper is organized as follows. Section II formulates the multi-class classification problem and presents the RNSC algorithm. Section III provides theoretical performance analysis. Numerical examples are provided in Section IV for handwritten digit datasets. Finally we conclude in Section V.

Remark: Throughout the paper, we use upper case bold letters for matrices, and lower case bold letters for vectors. Let $\|\mathbf{A}\|$, $\|\mathbf{A}\|_F$ denote the spectral norm and the Frobenius norm of $\mathbf{A}$ respectively. Let $\odot$ denote point-wise multiplication, $\mathrm{diag}(\mathbf{A})$ denote the diagonal matrix of $\mathbf{A}$, $\mathbf{I}$ denote the identity matrix, and $\mathbf{P_B}$ denote the orthogonal projection to the subspace spanned by the columns of $\mathbf{B}$.

II. Robust Nearest Subspace Classifier

A. Problem Formulation

Given a test sample $\mathbf{y} \in \mathbb{R}^M$, the purpose of multi-class classification is to assign $\mathbf{y}$ to one of the K classes. We assume there are n_i training samples from the ith class, and all training samples are stacked into a matrix as

$$\mathbf{X}_i = [\mathbf{x}_{i,1}, \cdots, \mathbf{x}_{i,n_i}] \in \mathbb{R}^{M \times n_i},$$

where $\mathbf{x}_{i,j} \in \mathbb{R}^M$ is the jth training sample in the ith class. Without loss of generality, we assume all samples are normalized, i.e. $\|\mathbf{x}_{i,j}\|_2 = 1$ and $\|\mathbf{y}\|_2 = 1$. The NSC [1] first calculates the distance from the test sample $\mathbf{y}$ to the ith class and measures the projection residual r_i from $\mathbf{y}$ to the orthogonal principal subspace $\mathbf{B}_i \in \mathbb{R}^{M \times k}$ of the training sets $\mathbf{X}_i$, which is spanned by the principal eigenvectors of $\mathbf{\Sigma}_i = \mathbf{X}_i\mathbf{X}_i^T$ for the ith class, given as

$$r_i = \|(\mathbf{I} - \mathbf{P}_{\mathbf{B}_i})\,\mathbf{y}\|_2 = \left\|\left(\mathbf{I} - \mathbf{B}_i\mathbf{B}_i^T\right)\mathbf{y}\right\|_2. \qquad (1)$$

The test sample $\mathbf{y}$ is then assigned to the class with the smallest residual among all classes, i.e.

$$i^* = \underset{i}{\operatorname{argmin}}\ r_i.$$

In this paper, both the training samples and the test samples suffer from the missing data problem, i.e. only a small fraction of the entries of $\mathbf{X}_i$'s and $\mathbf{y}$ are observed. We denote the partially observed test sample as

$$\mathbf{y}_\Omega = \mathbf{P}_\Omega \mathbf{y}, \tag{2}$$

where $\mathbf{P} = \mathbf{I}_\Omega \in \mathbb{R}^{m\times M}$ is a partial identity matrix where m rows are selected uniformly at random, denoted by the index set Ω. We denote the partially observed training matrix as

$$\mathbf{Z}_i = [\mathbf{z}_{i,1}, \cdots, \mathbf{z}_{i,n_i}] = \mathbf{P}_i \odot \mathbf{X}_i, \quad i = 1, \ldots, K, \tag{3}$$

where $\mathbf{P}_i = [p_i(k,j)] \in \{0,1\}^{M\times n_i}$ is a binary matrix where $p_i(k,j) = 1$ if the kth entry of the jth training sample in the ith class is observed, and $p_i(k,j) = 0$ if that entry is missing. We assume each entry is observed with probability $\delta \in [0,1]$ independently, and the goal is to design a robust version of the NSC in order to handle missing entries in the data.

B. Algorithm Details

Our robust nearest subspace classifier (RNSC) is proposed based on two modifications of the original NSC algorithm. The first change handles missing data in the training sets, where we use an unbiased estimator of $\mathbf{\Sigma}_i$ up to a scalar to calculate the principal subspaces from the training sets with missing entries, given as (5). The expectation of $\hat{\mathbf{\Sigma}}_i$ can be verified as $\mathbb{E}\hat{\mathbf{\Sigma}}_i = \delta^2\mathbf{\Sigma}_i$, however this estimator requires the knowledge of δ. In practice, we could consider an alternative estimator as

$$\tilde{\mathbf{\Sigma}}_i = \mathbf{Z}_i\mathbf{Z}_i^T \tag{4}$$

which is biased but doesn't require knowing δ. This is equivalent to setting $\delta = 1$ in (5), and equivalent to the original NSC by filling all missing entries as zero. In the numerical examples in Section 4, the performance of using $\tilde{\mathbf{\Sigma}}_i$ only degenerates a little compared with using $\hat{\mathbf{\Sigma}}_i$.

The second modification handles missing data in the test sample, where the residual to each class is calculated as (6), i.e. the distance between the test sample on the observed entries $\mathbf{y}_\Omega$ to the subspace spanned by $\hat{\mathbf{B}}_{i,\Omega} = \mathbf{I}_\Omega\hat{\mathbf{B}}_i$, obtained by restricting to the observed rows of the principal subspace $\hat{\mathbf{B}}_i$ extracted from $\hat{\mathbf{\Sigma}}_i$. Then the test sample is classified to the class with the smallest residual $\hat{r}_{i,\Omega}$. Algorithm 1 describes the details of the proposed RNSC algorithm.

III. Theoretical Analysis

In our theoretical analysis, we want to establish the relationship between the performance of RNSC using the unbiased estimator $\hat{\mathbf{\Sigma}}_i$ when missing data are present and the performance of the NSC when full data are available, by showing that with high probability, the residual $\hat{r}_{i,\Omega}$ computed from the RNSC for each class will be very close to the residual r_i computed from the NSC up to a scalar.

Algorithm 1 Robust Nearest Subspace Classifier (RNSC)

Input: training samples of the ith class $\mathbf{Z}_i$, $i = 1, \cdots, K$, the observation probability p, the test sample $\mathbf{y}_\Omega$;
Output: the label of the test sample $\mathbf{y}_\Omega$.

1: Compute the covariance matrix of each class using:

$$\hat{\mathbf{\Sigma}}_i = (\delta - 1)\,\mathrm{diag}\left(\mathbf{Z}_i\mathbf{Z}_i^T\right) + \mathbf{Z}_i\mathbf{Z}_i^T \tag{5}$$

2: Compute the principal subspace of $\hat{\mathbf{\Sigma}}_i$ of rank k as $\hat{\mathbf{B}}_i$;
3: Compute the distance from $\mathbf{y}_\Omega$ to each $\hat{\mathbf{B}}_i$ as

$$\hat{r}_{i,\Omega} = \left\|\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{B}}_{i,\Omega}}\right)\mathbf{y}_\Omega\right\|_2, \tag{6}$$

where $\mathbf{P}_{\hat{\mathbf{B}}_{i,\Omega}}$ is the orthogonal projection to the subspace spanned by $\hat{\mathbf{B}}_{i,\Omega}$ by restricting to the rows in Ω.
4: Claim the label of $\mathbf{y}_\Omega$ as

$$i^* = \underset{1\le i\le K}{\operatorname{argmin}}\,\hat{r}_{i,\Omega}. \tag{7}$$

A. Missing data in the training sets

We first describe the noncommutative Bernstein's inequality [12] in order to establish that $\|\hat{\mathbf{\Sigma}}_i - \delta^2\mathbf{\Sigma}_i\|$ is small with high probability.

Lemma 1: [12] Let $\mathbf{X}_1, \cdots, \mathbf{X}_L$ be independent zero-mean symmetric random matrices of dimension $M \times M$. Suppose $\sigma^2 = \sum_{k=1}^L \left\|\mathbb{E}[\mathbf{X}_k\mathbf{X}_k^T]\right\|$ and $\|\mathbf{X}_k\| \le B$ almost surely for all k. Then for any $0 < \tau < \sigma^2/B$,

$$\Pr\left[\left\|\sum_{k=1}^L \mathbf{X}_k\right\| > \tau\right] \le 2M\exp\left(-\frac{3\tau^2}{8\sigma^2}\right). \tag{8}$$

We next define the coherence of a subspace [13] as follows.

Definition 1: (Coherence of a subspace) Let $\mathbf{V}$ be a subspace of $\mathbb{R}^M$ of dimension k and $\mathbf{P}_\mathbf{V}$ be the orthogonal projection onto $\mathbf{V}$. Then the coherence of $\mathbf{V}$ is defined as

$$\mu(\mathbf{V}) = \frac{M}{k}\max_{1\le i\le M}\|\mathbf{P}_\mathbf{V}\mathbf{e}_i\|_2^2 \tag{9}$$

where $\{\mathbf{e}_i\}_{i=1}^M$ are standard basis vectors.

Note that for any subspace $1 \le \mu(\mathbf{V}) \le M/k$. We have the following theorem.

Theorem 1: Let $0 < \eta < 1$. Suppose

$$\delta > \sqrt{\frac{8}{3\|\mathbf{\Sigma}_i\|}\log\left(\frac{2M}{\eta}\right)}, \tag{10}$$

then with probability at least $1 - \eta$,

$$\left\|\hat{\mathbf{\Sigma}}_i - \delta^2\mathbf{\Sigma}_i\right\| \le \delta\sqrt{\frac{8\|\mathbf{\Sigma}_i\|}{3}\log\left(\frac{2M}{\eta}\right)}. \tag{11}$$

Proof: Define the matrix $\mathbf{Z}_{i,j} = \mathbf{z}_{i,j}\mathbf{z}_{i,j}^T$, $\mathbf{D}_{i,j} = \mathrm{diag}\left(\mathbf{z}_{i,j}\mathbf{z}_{i,j}^T\right)$ and $\mathbf{X}_{ij} = \mathbf{x}_{i,j}\mathbf{x}_{i,j}^T$, then

$$\hat{\mathbf{\Sigma}}_i - \delta^2\mathbf{\Sigma}_i = \sum_{j=1}^{n_i}((\delta-1)\mathbf{D}_{i,j} + \mathbf{Z}_{i,j} - \delta^2\mathbf{X}_{i,j}) \triangleq \sum_{j=1}^{n_i}\mathbf{V}_{i,j}.$$

where $\mathbf{V}_{i,j} = (\delta-1)\mathbf{D}_{i,j}+\mathbf{Z}_{i,j}-\delta^2\mathbf{X}_{i,j}$. It is straightforward that

$$\begin{aligned}\|\mathbf{V}_{ij}\| &\le (1-\delta)\|\mathbf{D}_{i,j}\| + \|\mathbf{Z}_{i,j}\| + \delta^2\|\mathbf{X}_{i,j}\| \\ &\le (1-\delta)+1+\delta^2 \le 2, \end{aligned} \tag{12}$$

where (12) follows from $\|\mathbf{X}_{i,j}\|_2 = \|\mathbf{x}_{i,j}\| = 1$, $\|\mathbf{Z}_{i,j}\|_2 = \|\mathbf{z}_{i,j}\| \le 1$. Now define the matrix $\mathbf{W}_{i,j}$ as

$$\begin{aligned}\mathbf{W}_{i,j} &= \mathbf{V}_{i,j}\mathbf{V}_{i,j} \\ &= (\delta-1)^2\mathbf{D}_{i,j}^2 + (\delta-1)\mathbf{D}_{i,j}(\mathbf{Z}_{i,j}-\delta^2\mathbf{X}_{i,j}) \\ &\quad + (\delta-1)(\mathbf{Z}_{i,j}-\delta^2\mathbf{X}_{i,j})\mathbf{D}_{i,j} + (\mathbf{Z}_{i,j}-\delta^2\mathbf{X}_{i,j})^2,\end{aligned}$$

where the diagonal entries of $\mathbb{E}[\mathbf{W}_{i,j}]$ is given by

$$\mathbb{E}[w_{i,j}(k,k)] = (\delta^3-\delta^2)x_{i,j}(k)^4 + (\delta^2-\delta^4)x_{i,j}(k)^2, \tag{13}$$

and the off-diagonal entries of $\mathbb{E}[\mathbf{W}_{i,j}]$ is given by

$$\mathbb{E}[w_{i,j}(k,\ell)] = (\delta^3-\delta^4)x_{i,j}(k)x_{i,j}(\ell), \tag{14}$$

where $x_{i,j}(k)$ and $z_{i,j}(k)$ denotes the kth entry of $\mathbf{x}_{i,j}$ and $\mathbf{z}_{i,j}$ respectively. Combining (13) and (14), we can rewrite $\mathbb{E}[\mathbf{W}_{ij}]$ as

$$\mathbb{E}[\mathbf{W}_{ij}] = (\delta^2-\delta^3)\left[\mathrm{diag}(\mathbf{X}_{i,j}) - \mathrm{diag}(\mathbf{X}_{i,j})^2\right] + (\delta^3-\delta^4)\mathbf{X}_{ij}.$$

We could bound $\sigma^2 = \left\|\sum_{j=1}^{n_i}\mathbb{E}[\mathbf{W}_{i,j}]\right\|$ as

$$\begin{aligned}\sigma^2 &\le (\delta^2-\delta^3)\left\|\sum_{j=1}^{n_i}\left[\mathrm{diag}(\mathbf{X}_{i,j}) - \mathrm{diag}(\mathbf{X}_{i,j})^2\right]\right\| \\ &\quad + (\delta^3-\delta^4)\left\|\sum_{i=1}^{n_i}\mathbf{X}_{i,j}\right\| \qquad (15) \\ &\le (\delta^2-\delta^3)\left\|\sum_{j=1}^{n_i}\mathrm{diag}(\mathbf{X}_{i,j})\right\| + (\delta^3-\delta^4)\left\|\sum_{i=1}^{n_i}\mathbf{X}_{i,j}\right\| \qquad (16) \\ &\le (\delta^2-\delta^3)\left\|\mathrm{diag}\left(\mathbf{\Sigma}_i\right)\right\| + (\delta^3-\delta^4)\left\|\mathbf{\Sigma}_i\right\| \qquad (17) \\ &\le (\delta^2-\delta^3)\|\mathbf{\Sigma}_i\| + (\delta^3-\delta^4)\left\|\mathbf{\Sigma}_i\right\| \qquad (18) \\ &\le \delta^2\|\mathbf{\Sigma}_i\|. \qquad (19)\end{aligned}$$

where (16) follows from the fact the diagonal entries of $\mathbf{X}_{i,j}$ is not greater than 1, and (17) follows by writing $\mathbf{\Sigma} = \sum_{j=1}^{n_i}\mathbf{X}_{i,j}$, (18) follows from the eigenvalue majorization.

Let $0 < \tau < \delta^2\|\mathbf{\Sigma}_i\|$, then following Lemma 1 we have

$$\Pr\left[\left\|\hat{\mathbf{\Sigma}}_i - \delta^2\mathbf{\Sigma}_i\right\| > \tau\right] \le 2M\exp\left(-\frac{3\tau^2}{8\delta^2\|\mathbf{\Sigma}_i\|}\right). \tag{20}$$

By letting $\tau = \delta\sqrt{\frac{8\|\mathbf{\Sigma}_i\|}{3}\log\left(\frac{2M}{\eta}\right)}$, we obtain (11). Since $\tau < \delta^2\|\mathbf{\Sigma}_i\|$, we have

$$\delta\sqrt{\frac{8\|\mathbf{\Sigma}_i\|}{3}\log\left(\frac{2M}{\eta}\right)} < \delta^2\|\mathbf{\Sigma}_i\|,$$

which gives (10). ∎

Theorem 1 does not make any assumptions on the training sets such as low rankness or incoherence conditions of the sample covariance matrix $\mathbf{\Sigma}_i$, which makes it highly versatile. Since $\|\mathbf{\Sigma}_i\| \le n_i$, we have the following corollary which provides an explicit bound on the number of training samples in each class.

Corollary 2: Let $0 < \eta < 1$. Suppose

$$\delta > \sqrt{\frac{8}{3n_i}\log\left(\frac{2M}{\eta}\right)}, \tag{21}$$

then with probability at least $1-\eta$,

$$\left\|\hat{\mathbf{\Sigma}}_i - \delta^2\mathbf{\Sigma}_i\right\| \le \delta\sqrt{\frac{8n_i}{3}\log\left(\frac{2M}{\eta}\right)}. \tag{22}$$

Denote the event that (11) happens by $\mathcal{G}$. Let $\mathbf{P}_{\hat{\mathbf{B}}_i}$ be the orthogonal projection to $\hat{\mathbf{B}}_i$, which is the principal subspace extracted from $\hat{\mathbf{\Sigma}}_i$, and define

$$\hat{r}_i = \left\|\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{B}}_i}\right)\mathbf{y}\right\|_2. \tag{23}$$

Our next step is to use the $\sin\mathbf{\Theta}$ Theorem of Davis and Kahan [14] in Lemma 2 to bound the distance between r_i and $\hat{r}_i$ under event $\mathcal{G}$.

Lemma 2: (Davis-Kahan) Let $\mathbf{A}$, $\tilde{\mathbf{A}}$ be $n\times n$ symmetric matrices, where $\tilde{\mathbf{A}}$ is a perturbed version of $\mathbf{A}$. Denote by λ_k the kth largest eigenvalue of $\mathbf{A}$ and $\mathbf{V}$ the eigenspace corresponding to the first k eigenvalues of $\mathbf{A}$. Denote by $\tilde{\sigma}_k$ and $\tilde{\mathbf{V}}$ the analogous quantities for $\tilde{\mathbf{A}}$. If the kth eigengap $\lambda_k - \lambda_{k+1} \ge \alpha$, then the distance between the two subspaces $\mathbf{V}$ and $\tilde{\mathbf{V}}$ is bounded by

$$\|\sin\mathbf{\Theta}\| \le \frac{\|\tilde{\mathbf{A}}-\mathbf{A}\|}{\alpha}, \tag{24}$$

where $\mathbf{\Theta} = [\theta_1,\cdots,\theta_k]$ is the canonical angles between the column space of $\mathbf{V}$ and $\tilde{\mathbf{V}}$.

Denote the normalized kth eigengap of $\mathbf{\Sigma}_i = \mathbf{X}_i\mathbf{X}_i^T$ as

$$\alpha_i = \frac{\lambda_k(\mathbf{\Sigma}_i) - \lambda_{k+1}(\mathbf{\Sigma}_i)}{n_i}. \tag{25}$$

Using the relationship in [14] that

$$\left\|\mathbf{P}_{\mathbf{B}_i} - \mathbf{P}_{\hat{\mathbf{B}}_i}\right\|_F = \sqrt{2}\left\|\sin\mathbf{\Theta}_i\right\|, \tag{26}$$

we have under the event $\mathcal{G}$,

$$\begin{aligned}\left|r_i - \hat{r}_i\right| &\le \|(\mathbf{P}_{\mathbf{B}_i} - \mathbf{P}_{\hat{\mathbf{B}}_i})\mathbf{y}\|_2 \\ &\le \|\mathbf{P}_{\mathbf{B}_i} - \mathbf{P}_{\hat{\mathbf{B}}_i}\|_F\|\mathbf{y}\|_2 = \sqrt{2}\|\sin\mathbf{\Theta}_i\| \qquad (27) \\ &\le \frac{\sqrt{2}}{n_i\delta^2\alpha_i}\|\hat{\mathbf{\Sigma}}_i - \delta^2\mathbf{\Sigma}_i\| \qquad (28) \\ &\le \frac{4}{\delta\alpha_i}\sqrt{\frac{2}{3n_i}\log\left(\frac{2M}{\eta}\right)}, \qquad (29)\end{aligned}$$

where (27) follows from $\|\mathbf{y}\|_2 = 1$ and (26), (29) follows from Lemma 2. It shows that given $\delta \gtrsim \mathcal{O}((\log M/n_i)^{1/2})$, as long as $n_i \gtrsim \mathcal{O}(\log M)$, the bound $|\hat{r}_i - r_i|$ is small with high probability.

B. Missing data in the test sample

Next we use the matched subspace detector developed in [11] to show the distance between $\hat{r}_i$ and $\hat{r}_{i,\Omega}$ is small as long as the number of observed entries in the test sample is about $m \gtrsim \mathcal{O}(k\mu(\hat{\mathbf{B}}_i)\log k)$. We first present the following lemma that slightly improves the lower bound in [11].

Lemma 3: ([11]) Let $0 < \epsilon < 1$ and $m = |\Omega| \geq \frac{8k}{3}\mu(\hat{\mathbf{B}}_i)\log(\frac{2k}{\epsilon})$. Then with probability at least $1-4\epsilon$,

$$(1-\beta)\hat{r}_i^2 \leq \frac{M}{m}\hat{r}_{i,\Omega}^2 \leq (1+\gamma)\hat{r}_i^2, \tag{30}$$

where

$$\gamma = \mu\left((\mathbf{I}-\mathbf{P}_{\hat{\mathbf{B}}_i})\mathbf{y}\right)\sqrt{\frac{2}{m}\log\left(\frac{1}{\epsilon}\right)}$$

and

$$\beta = \sqrt{\frac{8}{3m}\left(k\mu(\hat{\mathbf{B}}_i)+\mu\left((\mathbf{I}-\mathbf{P}_{\hat{\mathbf{B}}_i})\mathbf{y}\right)\right)\log\left(\frac{2(k+1)}{\epsilon}\right)}.$$

Proof: The RHS of (30) is the same as [11, Theorem 1]. For the lower bound, let $\mathbf{V} = [\hat{\mathbf{B}}_i, \mathbf{y}]$, and $\mathbf{V}_\Omega = [\hat{\mathbf{B}}_{i,\Omega}, \mathbf{y}_\Omega]$ be the subsampled matrix of $\mathbf{V}$ on the rows indexed by Ω. From [15, Lemma 5], we have that

$$\hat{r}_{i,\Omega}^2 \geq \lambda_{\min}(\mathbf{V}_\Omega^T\mathbf{V}_\Omega). \tag{31}$$

Furthermore, define the matrix $\mathbf{Q}$ as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{I} & -\frac{1}{\hat{r}_i}\hat{\mathbf{B}}_i^T\mathbf{y} \\ \mathbf{0} & \frac{1}{\hat{r}_i} \end{bmatrix}, \tag{32}$$

then we have $\mathbf{Q}^T\mathbf{V}^T\mathbf{V}\mathbf{Q} = \mathbf{I}$. Define

$$\tilde{\mathbf{V}} = \mathbf{V}\mathbf{Q} = \left[\hat{\mathbf{B}}_i \quad \frac{1}{\hat{r}_i}(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{B}}_i})\mathbf{y}\right].$$

The coherence of $\tilde{\mathbf{V}}$ satisfies

$$\mu(\tilde{\mathbf{V}}) \leq \frac{k\mu(\hat{\mathbf{B}}_i)+\mu\left((\mathbf{I}-\mathbf{P}_{\hat{\mathbf{B}}_i})\mathbf{y}\right)}{k+1}.$$

Using similar arguments as [11, Lemma 3], we have

$$\left\|\mathbf{Q}^T\mathbf{V}_\Omega^T\mathbf{V}_\Omega\mathbf{Q} - \frac{m}{M}\mathbf{I}\right\| \leq \frac{m}{M}\beta \tag{33}$$

Therefore, $\lambda_{\min}(\mathbf{V}_\Omega^T\mathbf{V}_\Omega) \geq \frac{m}{M}(1-\beta)\lambda_{\max}(\mathbf{Q}^T\mathbf{Q})^{-1} = \frac{m}{M}(1-\beta)\hat{r}_i^2$. Plugging this in (31) yields (30). ∎

Denote the event that (30) happens by $\mathcal{H}$. It is worth noting that we can bound the number of observed entries m using $\mu(\mathbf{B}_i)$ based on the following bound between the coherence of $\hat{\mathbf{B}}_i$ and $\mathbf{B}_i$ under event $\mathcal{G}$:

$$\begin{aligned}\mu(\hat{\mathbf{B}}_i)^{\frac{1}{2}} &\leq \sqrt{\frac{M}{k}}\left\|\mathbf{P}_{\hat{\mathbf{B}}_i}-\mathbf{P}_{\mathbf{B}_i}\right\|_F\|\mathbf{e}_i\|_2+\mu(\mathbf{B}_i)^{\frac{1}{2}} \\ &\leq \frac{4}{\delta\alpha_i}\sqrt{\frac{2M}{3n_ik}\log\left(\frac{2M}{\eta}\right)}+\mu(\mathbf{B}_i)^{\frac{1}{2}},\end{aligned} \tag{34}$$

Similarly, we can lower bound $\mu(\hat{\mathbf{B}}_i)$ as

$$\mu(\hat{\mathbf{B}}_i)^{\frac{1}{2}} \geq \mu(\mathbf{B}_i)^{\frac{1}{2}} - \frac{4}{\delta\alpha_i}\sqrt{\frac{2M}{3n_ik}\log\left(\frac{2M}{\eta}\right)}. \tag{35}$$

C. Connecting the Pieces

The main theorem is to connect the residual r_i when no missing data are present for each class is close to $\hat{r}_{i,\Omega}$ when missing data are present by combining (29) and (30) under the event $\mathcal{G}\cap\mathcal{H}$. We have the following theorem.

Theorem 3: Let $0 < \eta, \epsilon < 1$. Given δ satisfies (10) and $m \geq \frac{8k}{3}\mu(\hat{\mathbf{B}}_i)\log(\frac{2k}{\epsilon})$, with probability at least $1-\eta-4\epsilon$,

$$\sqrt{\frac{M}{m}}\hat{r}_{i,\Omega} \leq \sqrt{1+\gamma}\left[r_i + \frac{4}{\delta\alpha_i}\sqrt{\frac{2}{3n_i}\log\left(\frac{2M}{\eta}\right)}\right],$$

$$\sqrt{\frac{M}{m}}\hat{r}_{i,\Omega} \geq \sqrt{1-\beta}\left[r_i - \frac{4}{\delta\alpha_i}\sqrt{\frac{2}{3n_i}\log\left(\frac{2M}{\eta}\right)}\right].$$

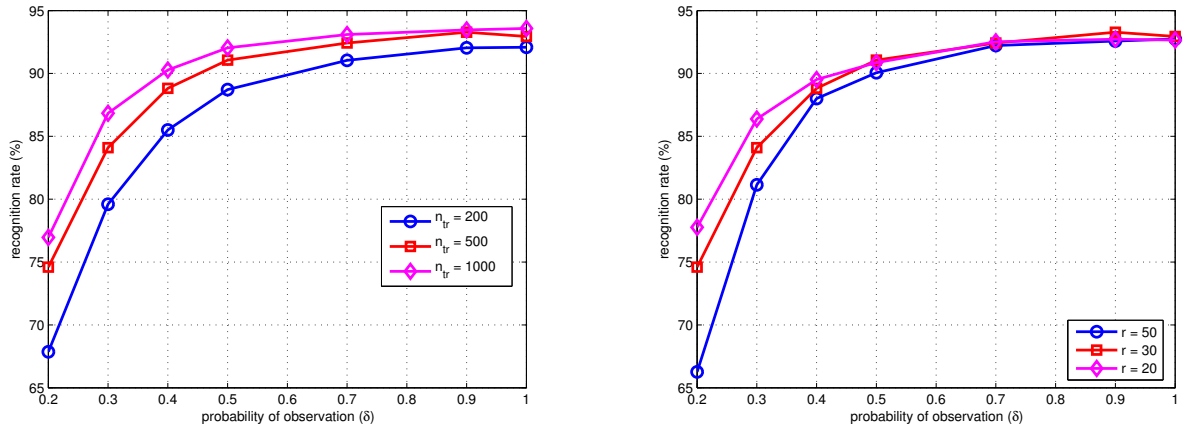
Our theorem indicates that as long as the percentage of observed entries in the training set is $\delta \gtrsim \mathcal{O}((\log M/n_i)^{1/2})$, and the number of observed entries in the test sample is $m \gtrsim \mathcal{O}(k\mu(\mathbf{B}_i)\log k)$ for all classes, the performance of the RNSC is close to the performance of the original NSC when full data is available.

IV. Numerical Examples

In this section, we examined the proposed RNSC algorithms for digit recognition, when both the training samples and testing samples have a significant percentage of missing entries. We use the MNIST Handwritten Digits database [16], which include about 6000 training samples and 1000 test samples per digit in the data set. Each sample is an 8-bit gray-scale image of "0" through "9" with dimension $M = 28\times 28 = 784$.

We randomly choose $n_{tr} = 1000$ samples for training and $n_{te} = 500$ samples for testing per digit. We assume each pixel in both the training and the testing sample images is observed with the same probability $\delta \in (0,1]$ because they may experience the same impairments in the system. Fig. 2 shows the recognition rate versus δ, when a principal subspace of rank $k = 30$ is extracted using the unbiased estimator $\hat{\mathbf{\Sigma}}_i$ and the biased estimator $\tilde{\mathbf{\Sigma}}_i$ for each class from the training data. With observing only 20% of the whole data, it still achieves a recognition rate of 77% for both estimators. With 50% of the data, the recognition rate is within 2% degradation of the case when full data is available. It is also worth noting that the performance of using the biased estimator is only slightly worse (within 1% percent) than that of using the unbiased estimator, but does not require the knowledge of δ.

Fig. 1 (a) shows the recognition rate with respect to δ for the unbiased estimator with different training size per digit when the rank of the principal subspace is fixed as $k = 30$. The recognition rate increases as the training size increases as the estimator is expected to perform better. Fig. 1 (b) shows the recognition rate with respect to δ for varied rank of the principal subspace $k = 20, 30, 50$ when the training size is $n_{tr} = 500$ per digit. When δ is relatively small, the difference between recognition rates for various ranks is bigger compared when δ is large, and rank $k = 20$ performs best. This may indicate that more gain can be achieved by selecting the optimal subspace rank when there are missing data.



(a) (b)

Fig. 1. Recognition rate with respect to probability of observation, for the unbiased covariance estimator with (a) different training size per class for $r = 30$; (b) different principal subspace rank with $n_{tr} = 500$ per class.

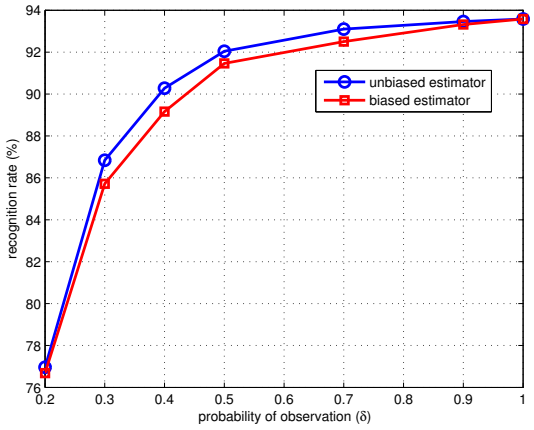


Fig. 2. Recognition rate with respect to probability of observation of the proposed RNSC, for the unbiased and biased estimators with $r = 30$.

V. Conclusions

In this paper, we proposed and analyzed a robust version of the nearest subspace classifier, dubbed robust nearest subspace classifier (RNSC), when only a small amount of entries are observed in both the training samples and testing samples. We show the scaling between the number of training samples, the amount of missing data when its performance is comparable to the nearest subspace classifier when no missing data are present if the corresponding spectral gap is not too small. We show that our algorithm achieves performance close to the scenario when no missing data are present on real-world data sets.

Acknowledgements

This work was partially supported by a grant from the Simons Foundation.

References

[1] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. on PAMI*, vol. 27, no. 5, pp. 684–698, 2005.

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on PAMI*, vol. 31, no. 2, 2009.

[3] E. J. Candés, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[4] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 1289–1306, Feb. 2006.

[5] Y. Chi and F. Porikli, "Connecting the dots: From nearest subspace to collaborative representation," in *CVPR*, 2012.

[6] K. Lounici, "High-dimensional covariance matrix estimation with missing observations," *Arxiv*, May 2012.

[7] Y. Chi, Y. C. Eldar, and R. Calderbank, "Petrels: Subspace estimation and tracking from partial observations," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3301–3304.

[8] ——, "Petrels: Parallel subspace estimation and tracking by recursive least squares from partial observations," *IEEE Trans. on Signal Processing*, vol. 61, pp. 5947 – 5959, 2013.

[9] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," *Proc. Allerton*, 2010.

[10] Y. Xie, J. Huang, and R. Willett, "Multiscale online tracking of manifolds," in *Statistical Signal Processing Workshop (SSP), 2012 IEEE*. IEEE, 2012, pp. 620–623.

[11] L. Balzano, B. Recht, and R. Nowak, "High-dimensional matched subspace detection when data are missing," in *Proc. ISIT*, June 2010.

[12] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[13] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2053–2080, 2009.

[14] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*. Academic Press, 1990.

[15] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *Information Theory, IEEE Transactions on*, vol. 57, no. 7, pp. 4680–4688, 2011.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.