SUBSPACE DETECTION IN A KERNEL SPACE: THE MISSING DATA CASE

Tong Wu and Waheed U. Bajwa

Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey

ABSTRACT

This paper studies the problem of matched subspace detection in high-dimensional feature space where the signal in the input space is partially observed. We present a test statistic for our detection problem using kernel functions and provide kernel function value estimators with missing data for different kernels. The test statistic can be calculated approximately with estimated kernel function values. We also give theoretical results regarding the kernel function value and test statistic estimation. Numerical experiments involving both Gaussian and polynomial kernels show the efficacy of the proposed kernel function value estimator and resulting subspace detector.

***Index Terms*—**Kernel methods, missing data, subspace detector

1. INTRODUCTION

Modeling high-dimensional signals as living in a low-dimensional subspace is one of the most widely used signal models for a collection of high-dimensional data. Matched subspace detection, where the goal is to test whether a signal lies in a known low-dimensional subspace, arises in the literature in many guises such as hyperspectral imaging [1], radar signal processing [2] and compressive sensing [3]. The problem can be formally described as follows. Suppose we have an observation signal $x = s + w \in \mathbb{R}^n$ where s denotes the signal of interest and w is a noisy term with known distribution. The decision problem is to distinguish between two possible hypotheses: $\mathcal{H}_0 : s \in \mathcal{S}$ and $\mathcal{H}_1 : s \notin \mathcal{S}$, where $\mathcal{S}$ is a given subspace of $\mathbb{R}^n$. There have been some efforts in the community to develop optimum matched subspace detectors [4, 5] and in general, these detectors always depend on the amount of energy of x in the subspace $\mathcal{S}$. In many cases, linear subspace model is sufficient to represent the geometry underlying the signals; however, in many emerging applications, ensembles cannot be well approximated by linear models in the ambient space. In the past two decades, kernel methods have been widely studied [6], the basic idea underlying which is to introduce a nonlinear mapping Φ from the data space to a high-dimensional (possibly infinite-dimensional) feature space $\mathcal{F}$ such that the mapped "images" can be modeled as linear in $\mathcal{F}$. Recently, a nonlinear, kernel version of the matched subspace detector has been proposed in [1] and it was shown to be successful for hyperspectral target detection in the presence of noise and background.

In this paper, we focus on the problem of kernel matched subspace detection in the presence of missing data. The setup here corresponds to the situation when we only observe $x = s + w$ in the input space at certain locations $\Omega \subset \{1, 2, \ldots, n\}$. Based on these observed entries and a given r-dimensional subspace $\mathcal{S} \subset \mathcal{F}$ ($r \ll n$), we wish to determine whether $\Phi(s) \in \mathcal{S}$. This problem is motivated because the data we are processing are often corrupted due to sensor failures, improper data collection procedures and noisy measurements (e.g., respondent's refusal to answer questions related to private subjects such as education and income). The matched subspace detector under linear model with missing data has been well studied in [7]. However, due to the lack of knowledge about the nonlinear mapping function and the very high dimensionality of the feature space, the problem formulation and treatment in here is different from that in [7].

The main idea of our work is as follows. Based on the idea described in [1], we first derive the test statistic expression with kernel representations for our detection problem in the presence of complete test data. The main challenge for the detection problem with missing test data is to calculate the test statistic with only partially observed input data. We address this challenge by proposing estimators of the kernel function values based on missing data under different types of kernels. Our results show that under some mild conditions, we can reliably create an estimator of the function value and it is possible to perform kernel matched subspace detection from very few entries. In order to demonstrate effectiveness of kernel function value estimation and subspace detection, we carry out numerical experiments via two commonly used kernels: Gaussian kernel and polynomial kernel. Results of these experiments show the effectiveness of our proposed methods and robustness to noise of the detection strategy.

2. PRELIMINARIES

2.1. Background and Problem Formulation

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be an n-dimensional input space and $\Phi : \mathcal{X} \to \mathcal{F}$ be a nonlinear mapping from $\mathcal{X}$ to a high-dimensional (possibly infinite-dimensional) feature space. In this case, every data point $x \in \mathcal{X}$ has an "image" $\Phi(x)$ in $\mathcal{F}$. However, directly working with $\Phi(x)$ increases the computational complexity and memory requirements tremendously because of the high dimensionality of the feature space. In order to circumvent this, kernel methods are used instead, which allow us to compute the dot products in $\mathcal{F}$ using input vectors in $\mathcal{X}$ instead of mapping them into $\mathcal{F}$ [6]. To be specific, a *kernel* function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ satisfies $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ for all $x, x' \in \mathcal{X}$. Compared with linear methods, kernel methods can explore nonlinear data structures. Besides, kernels can often be computed efficiently, even in very high-dimensional ambient spaces.

Now assume we are given a vector x_Ω of dimension $m = |\Omega|$ corresponding to the entries of a noisy test sample $x = s + w \in \mathbb{R}^n$ indexed by Ω, where s denotes a signal and w is additive white Gaussian noise with variance σ^2. We are interested in determining whether $\Phi(s)$ belongs to an r-dimensional subspace $\mathcal{S} \subset \mathcal{F}$ or not based only on these m known coordinates. In here, we assume that the subspace dimension r is known *a priori* and $r \ll n$. This problem can be described as the composite hypothesis test:

$$\mathcal{H}_0 : \Phi(s) \in \mathcal{S} \quad \text{vs.} \quad \mathcal{H}_1 : \Phi(s) \notin \mathcal{S}. \tag{1}$$

In order to learn $\mathcal{S}$, we assume access to a collection of N "complete" noiseless training data points $X = \{x_p \in \mathbb{R}^n\}_{p=1}^N$ where all

This work is supported in part by an ARL Robotics CTA subaward.

$\Phi(x_p)$'s correspond to samples drawn from $\mathcal{S}$. We assume without loss of generality that all these training data x_p's have been normalized to unit ℓ_2 norms. We first review a method for learning the subspace in the feature space with training samples, termed kernel principle component analysis (PCA) algorithm. Then we will discuss the kernel matched subspace detector for the case of "complete" test data in Section 2.3.

2.2. Review of Kernel PCA

The subspace learning method we use in this paper is the well-known kernel PCA algorithm [6]. Let $\Phi(X) = [\Phi(x_1), \ldots, \Phi(x_N)]$ denote the matrix whose columns are "images" of the training data in the feature space. First of all, the Φ-mapped data in $\mathcal{F}$ need to be centered. We denote the mean of the Φ-mapped "images" by $\bar{\Phi} = \frac{1}{N}\sum_{p=1}^{N}\Phi(x_p)$, resulting in N centered training samples $\widetilde{\Phi}(x_p) = \Phi(x_p) - \bar{\Phi}$, $p = 1, \ldots, N$. We denote by $K \in \mathbb{R}^{N\times N}$ the kernel matrix with $K_{pq} = \langle \Phi(x_p), \Phi(x_q)\rangle = k(x_p, x_q)$. Therefore the centered kernel matrix is given by

$$\widetilde{K} = K - HK - KH + HKH, \tag{2}$$

where H is an $N \times N$ matrix with all elements $\frac{1}{N}$. Kernel PCA involves performing the eigen decomposition $\widetilde{K} = U\Lambda U^T$ where $U = [u_1, \ldots, u_N]$ is the matrix containing the eigenvectors and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ contains the corresponding eigenvalues in descending order. Note that any eigenvector of the covariance matrix with nonzero eigenvalue must lie in the span of the Φ-mapped "images," i.e., the k-th eigenvector v_k ($\lambda_k \neq 0$) can be written as [6]

$$v_k = \frac{1}{\sqrt{\lambda_k}}\widetilde{\Phi}(X)u_k \stackrel{def}{=} \widetilde{\Phi}(X)e_k, \tag{3}$$

where $\widetilde{\Phi}(X) = [\widetilde{\Phi}(x_1), \ldots, \widetilde{\Phi}(x_N)]$. Therefore the r-dimensional subspace $\mathcal{S}$ can be represented by its orthonormal basis $V = [v_1, \ldots, v_r] = \widetilde{\Phi}(X)E$ where $E = [e_1, \ldots, e_r]$ with $e_k = \frac{u_k}{\sqrt{\lambda_k}}$.

2.3. Kernel Matched Subspace Detector

In order to introduce the detection method that will be used afterward, let's first consider the problem of kernel matched subspace detection when the test vector $x \in \mathbb{R}^n$ is complete. The authors in [1] developed a related framework for hyperspectral target detection by addressing the following detection problem:

$$\mathcal{H}_0 : \Phi(x) = B_\Phi\zeta_\Phi + n_\Phi \text{ vs. } \mathcal{H}_1 : \Phi(x) = B_\Phi\zeta_\Phi + T_\Phi\theta_\Phi + n_\Phi,$$

where B_Φ and T_Φ are bases of the background subspace $\mathcal{S}_1$ and target subspace $\mathcal{S}_2$, ζ_Φ and θ_Φ are the subspace coefficients, and n_Φ represents Gaussian random noise in the feature space. The subspaces $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{F}$ can be learned from their corresponding training data. The test statistic in this case depends on (i) the energy of $\Phi(x)$ outside the union-of-subspaces $\mathcal{S}_1 \cup \mathcal{S}_2$ and (ii) the energy of $\Phi(x)$ outside subspace $\mathcal{S}_1$. While we have a slightly different problem setup, we can leverage the ideas of [1] and carry out the subspace detection in the feature space by measuring the energy of $\Phi(x)$ outside the subspace $\mathcal{S}$. Mathematically, we can write it as $\|\Phi(x) - P_\mathcal{S}\Phi(x)\|_2^2 = \|\widetilde{\Phi}(x) - P_\mathcal{S}\widetilde{\Phi}(x)\|_2^2$, where $P_\mathcal{S} = VV^T$ denotes the projection operator onto $\mathcal{S}$ and $\widetilde{\Phi}(x) = \Phi(x) - \bar{\Phi}$. It follows that the decision strategy can be written as

$$\begin{aligned} T(x) &= \|\widetilde{\Phi}(x) - P_\mathcal{S}\widetilde{\Phi}(x)\|_2^2 = \|\widetilde{\Phi}(x)\|_2^2 - \|V^T\widetilde{\Phi}(x)\|_2^2 \\ &= \widetilde{k}(x,x) - \|E^T\widetilde{\Phi}(X)^T\widetilde{\Phi}(x)\|_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta, \end{aligned} \tag{4}$$

where $\widetilde{k}(x,x) = \langle\widetilde{\Phi}(x), \widetilde{\Phi}(x)\rangle = k(x,x) - \frac{2}{N}\mathbf{1}_N^T\boldsymbol{k}_x + \frac{1}{N^2}\mathbf{1}_N^T K\mathbf{1}_N$. Here $\mathbf{1}_N = [1, 1, \ldots, 1]^T$ is an N-dimensional vector and $\boldsymbol{k}_x = [k(x, x_1), \ldots, k(x, x_N)]^T$. Denoting by $\widetilde{\boldsymbol{k}}_x = \widetilde{\Phi}(X)^T\widetilde{\Phi}(x) = \boldsymbol{k}_x - H\boldsymbol{k}_x - \frac{1}{N}K\mathbf{1}_N + \frac{1}{N}HK\mathbf{1}_N$, we see that the test statistic (4) can be expressed as

$$T(x) = \widetilde{k}(x,x) - \|E^T\widetilde{\boldsymbol{k}}_x\|_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta. \tag{5}$$

Now we are ready to consider the detection problem described in (1), where we only observe elements of x indexed by Ω. It can be seen from (5) that the computation of the kernel function $k(x, x')$ plays a key role in computing $T(x)$, where x' denotes any training signal or a complete test signal. We need to compute an estimate of the test statistic $T(x)$ in the presence of missing test data. The challenge in this regard is to find a method which can give us reliable estimates of the kernel function values based on x_Ω and x_p's.

3. ESTIMATION OF KERNEL FUNCTION VALUE FROM INCOMPLETE DATA

In this section we present our proposed estimator for $k(x, x')$. One naïve approach to estimating $k(x, x')$ is by zero-filling the incomplete vector x_Ω. However, this approach is flawed and the resulting kernel function estimation can deviate from the true value largely. A better approach is to find a proxy function $g(\cdot,\cdot)$ such that $g(x_\Omega, x'_\Omega) \approx k(x, x')$, where x'_Ω is a vector containing the elements of x' indexed by Ω. More specifically, we wish to estimate $k(x, x')$ with entries of x and x' corresponding to Ω only. In the following, we assume indices of the observed entries, Ω, of x are drawn uniformly at random with replacement from $\{1, \ldots, n\}$. Nonetheless, the results reported in here can be easily extended to the case of Ω without replacement (see, e.g., [8, Lemma 1]). To find the proxy function $g(\cdot,\cdot)$, we start by considering the relationship between x_Ω, x'_Ω and x, x' underlying different types of kernel functions.

We first consider isotropic kernels of the form $k(x, x') = \kappa(\|x - x'\|_2^2)$. Let $\gamma = x - x'$ and $\gamma_\Omega = x_\Omega - x'_\Omega \in \mathbb{R}^m$. For a vector γ, the authors in [7] derived the coherence of a subspace spanned by γ to be $\mu(\gamma) = \frac{n\|\gamma\|_\infty^2}{\|\gamma\|_2^2}$ and it is clear that $1 \leq \mu(\gamma) \leq n$. They also have shown in [7] that $\|\gamma_\Omega\|_2^2$ is very close to $\frac{m}{n}\|\gamma\|_2^2$ with high probability. The following Corollary is adapted from [7, Lemma 1] by plugging in the definitions of γ and γ_Ω.

Corollary 1. *Let $\delta > 0$ and $\alpha = \sqrt{\frac{2\mu(x-x')^2}{m}\log(\frac{1}{\delta})}$, then with probability at least $1 - 2\delta$,*

$$(1-\alpha)\|x - x'\|_2^2 \leq \frac{n}{m}\|x_\Omega - x'_\Omega\|_2^2 \leq (1+\alpha)\|x - x'\|_2^2.$$

With this simple relationship, we can replace the distance term $\|x - x'\|_2^2$ in the corresponding kernel function by $\frac{n}{m}\|x_\Omega - x'_\Omega\|_2^2$ and yield an estimate. For example, for the Gaussian kernel $k(x, x') = \exp(\frac{-\|x-x'\|_2^2}{c})$ where $c > 0$ is a constant, we can set $g(x_\Omega, x'_\Omega) = \exp(\frac{-n\|x_\Omega - x'_\Omega\|_2^2}{mc})$. This then leads to the following bounds for the Gaussian kernel value estimation.

Theorem 1. *Let $\delta > 0$ and $\alpha = \sqrt{\frac{2\mu(x-x')^2}{m}\log(\frac{1}{\delta})}$, then for a Gaussian kernel $k(x, x')$, with probability at least $1 - 2\delta$,*

$$g(x_\Omega, x'_\Omega)^{\frac{1}{1-\alpha}} \leq k(x, x') \leq g(x_\Omega, x'_\Omega)^{\frac{1}{1+\alpha}}. \tag{6}$$

We should also note that $g(x_\Omega, x_\Omega) = k(x,x) = 1$ for Gaussian kernels. Similarly, for dot product kernels of the form $k(x,x') = \kappa(\langle x, x'\rangle)$, we need to find the relationship between $\langle x, x'\rangle$ and $\langle x_\Omega, x'_\Omega\rangle$. Let $z = x \circ x'$ be the coordinate-wise product of x and x' and $z_\Omega \in \mathbb{R}^m$ be a subvector containing the entries of z indexed by Ω. In other words, $\langle x, x'\rangle$ and $\langle x_\Omega, x'_\Omega\rangle$ are summations of all the entries of z and z_Ω respectively. We have the following Lemma which describes the deviation of the estimated inner product.

Lemma 1. *Let $\delta > 0$ and $\epsilon = \sqrt{\frac{2n^2\|x\circ x'\|_\infty^2}{m}\log(\frac{1}{\delta})}$, then with probability at least $1-2\delta$,*

$$\langle x, x'\rangle - \epsilon \le \frac{n}{m}\langle x_\Omega, x'_\Omega\rangle \le \langle x, x'\rangle + \epsilon. \tag{7}$$

Proof. First, it is clear that $\langle x, x'\rangle = \sum_{j=1}^n z_j$ and $\langle x_\Omega, x'_\Omega\rangle = \sum_{i=1}^m z_{\Omega_i}$. Let $f(Y_1,\ldots,Y_m) = \sum_{i=1}^m Y_i$ and $Y_i = z_{\Omega_i}$. Note that the value of f is the sum of m random variables drawn uniformly from a set $\{z_1,\ldots,z_n\}$. We assume that these m variables are sampled with replacement, hence they are independent and we have $\mathbb{E}[\sum_{i=1}^m Y_i] = \mathbb{E}[\sum_{i=1}^m z_{\Omega_i}] = \frac{m}{n}\sum_{j=1}^n z_j$. Since $|z_{\Omega_i}| \le \|z\|_\infty$ for all i, if we replace any sample value Y_ℓ with $\hat{Y}_\ell$, we have

$$|\sum_{i=1}^m Y_i - \sum_{i\neq\ell} Y_i - \hat{Y}_\ell| = |Y_\ell - \hat{Y}_\ell| \le 2\|z\|_\infty.$$

Therefore, McDiarmid's Inequality [9] shows that for $\epsilon > 0$,

$$\mathbb{P}[|\sum_{i=1}^m Y_i - \frac{m}{n}\sum_{j=1}^n z_j| \ge \frac{m}{n}\epsilon] \le 2\exp(\frac{-m\epsilon^2}{2n^2\|z\|_\infty^2}),$$

or equivalently,

$$\mathbb{P}[\sum_{j=1}^n z_j - \epsilon \le \frac{n}{m}\sum_{i=1}^m Y_i \le \sum_{j=1}^n z_j + \epsilon] \ge 1 - 2\exp(\frac{-m\epsilon^2}{2n^2\|z\|_\infty^2}).$$

Plugging the definition of ϵ yields the desired result. □

Lemma 1 states that $\langle x_\Omega, x'_\Omega\rangle$ is close to $\frac{m}{n}\langle x, x'\rangle$ with high probability. We once again use this relationship and give an estimate of the corresponding kernel function value. For example, for the polynomial kernel $k(x,x') = (\langle x, x'\rangle + c)^d$ with degree $d > 0$ and offset $c \ge 0$, the proxy function will be $g(x_\Omega, x'_\Omega) = (\frac{n}{m}\langle x_\Omega, x'_\Omega\rangle + c)^d$. To analyze the bounds on the function value estimation, notice that if (7) holds and d is odd, then

$$(\langle x, x'\rangle - \epsilon + c)^d \le (\frac{n}{m}\langle x_\Omega, x'_\Omega\rangle + c)^d \le (\langle x, x'\rangle + \epsilon + c)^d$$

always holds. But the above cannot be guaranteed to hold for an even d. We obtain the theorem below for polynomial kernels.

Theorem 2. *Let $\delta > 0$ and $\epsilon = \sqrt{\frac{2n^2\|x\circ x'\|_\infty^2}{m}\log(\frac{1}{\delta})}$, then for a polynomial kernel $k(x,x')$ with an odd degree d, with probability at least $1-2\delta$,*

$$(g(x_\Omega, x'_\Omega)^{\frac{1}{d}} - \epsilon)^d \le k(x,x') \le (g(x_\Omega, x'_\Omega)^{\frac{1}{d}} + \epsilon)^d. \tag{8}$$

We conclude this section by giving some intuition for the bounds in Theorem 1 and 2. Without loss of generality, we consider the case when both x and x' are unit ℓ_2 norm vectors. From the bounds in both cases, we can see that the parameters α and ϵ depend on $\sqrt{\log(\frac{1}{\delta})}$ and these two parameters increase when δ gets very small. Thus we need a larger m to achieve the same result as for a larger δ. Theorem 1 further tells us that as $\mu(x-x')$ increases, it will require an increase in m to get the same bound. For the polynomial kernel function value estimation, we can also estimate how many samples of x must be observed so that we are confident about our estimation. An example case is when both $\|x\|_\infty$ and $\|x'\|_\infty$ are close to $\frac{1}{\sqrt{n}}$, resulting in $\|z\|_\infty \approx \frac{1}{n}$ and thus $g(x_\Omega, x'_\Omega)$ is very close to $k(x,x')$ with high probability once $m \sim O(1)$. But as $\|z\|_\infty$ increases, we should have more samples to achieve reliable estimation. For example, if both $\|x\|_\infty$ and $\|x'\|_\infty$ are close to $n^{-\frac{1}{4}}$, then $\|z\|_\infty \approx n^{-\frac{1}{2}}$, in which case we must have at least $m \sim O(\sqrt{n})$ elements of x to estimate $(\langle x, x'\rangle + c)^d$ in a robust manner.

4. MATCHED SUBSPACE DETECTION

In this section, we are interested in understanding the behavior of the estimated kernel function value for subspace detection. To evaluate the performance of the matched subspace detection, the false alarm and detection probabilities based on test statistic described in (5) can be written as $P_{FA} = \mathbb{P}(T(x) > \eta|\mathcal{H}_0)$ and $P_D = \mathbb{P}(T(x) > \eta|\mathcal{H}_1)$. The optimum detector is usually developed from two aspects: either (i) by maximizing the detection probability for a given false alarm probability or (ii) by minimizing the probability of false alarm for a given detection probability. To bridge the chasm between the detection performance in the presence of "complete" and "missing" test data, we need to analyze the error of the estimated $T(x)$. The following two corollaries show the bounds of $T(x)$ for Gaussian and polynomial kernels in terms of the estimated function values.

Corollary 2. *Define $\alpha_s = \max_p \sqrt{\frac{2\mu(x-x_p)^2}{m}\log(\frac{1}{\delta})}$ for a fixed $\delta > 0$. Then for a Gaussian kernel $k(x,x')$,*

$$T_{GauL}(x) \le T(x) \le T_{GauU}(x) \tag{9}$$

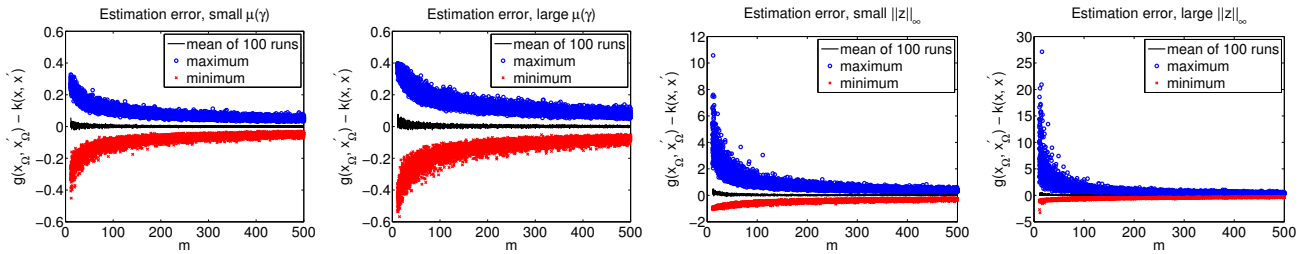
holds with probability at least $1-2N\delta$, where $T_{GauU}(x) = 1 + \frac{1}{N^2}\mathbf{1}_N^T K\mathbf{1}_N - \frac{2}{N}\sum_{p=1}^N g(x_\Omega, x_{p_\Omega})^{\frac{1}{1-\alpha_s}}$ and $T_{GauL}(x) = 1 - \frac{2}{N}\sum_{p=1}^N g(x_\Omega, x_{p_\Omega})^{\frac{1}{1+\alpha_s}} + \frac{1}{N^2}\mathbf{1}_N^T K\mathbf{1}_N - \Big(\|\frac{1}{N}E^T(H - I)K\mathbf{1}_N\|_2 + \|E^T(I-H)\|_2\sqrt{\sum_{p=1}^N g(x_\Omega, x_{p_\Omega})^{\frac{2}{1+\alpha_s}}}\Big)^2$.

Corollary 3. *Define $\epsilon_s = \sqrt{\frac{2n^2\max\{\|x\circ x\|_\infty^2, \max_p\|x\circ x_p\|_\infty^2\}}{m}\log(\frac{1}{\delta})}$ for a fixed $\delta > 0$. Then for a polynomial kernel $k(x,x')$ with an odd degree d,*

$$T_{polyL}(x) \le T(x) \le T_{polyU}(x) \tag{10}$$

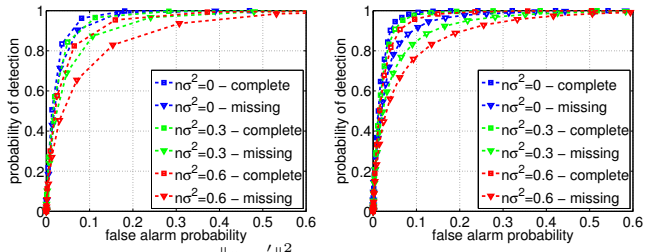
holds with probability at least $1-2(N+1)\delta$, where $T_{polyU}(x) = (g(x_\Omega, x_\Omega)^{\frac{1}{d}} + \epsilon_s)^d + \frac{1}{N^2}\mathbf{1}_N^T K\mathbf{1}_N - \frac{2}{N}\sum_{p=1}^N (g(x_\Omega, x_{p_\Omega})^{\frac{1}{d}} - \epsilon_s)^d$ and $T_{polyL}(x) = (g(x_\Omega, x_\Omega)^{\frac{1}{d}} - \epsilon_s)^d - \frac{2}{N}\sum_{p=1}^N (g(x_\Omega, x_{p_\Omega})^{\frac{1}{d}} + \epsilon_s)^d + \frac{1}{N^2}\mathbf{1}_N^T K\mathbf{1}_N - \Big(\|\frac{1}{N}E^T(H - I)K\mathbf{1}_N\|_2 + \|E^T(I-H)\|_2\sqrt{\sum_{p=1}^N \max\{(g(x_\Omega, x_{p_\Omega})^{\frac{1}{d}} \pm \epsilon_s)^{2d}\}}\Big)^2$.

We discuss the performance of the estimated $T(x)$ based on the two ideas mentioned above: (i) maximize the detection probability and (ii) minimize the false alarm rate. If the goal focuses on the probability of detection and $P_D = \mathbb{P}(T(x) > \eta_D|\mathcal{H}_1) = \tau_D$ in the complete data setting, we can use the same η_D as the test threshold and compute $T_{GauU}(x)$ or $T_{polyU}(x)$ as the test statistic in the missing data setting. This strategy admits high probability of



(a) $\mu(\gamma) \approx 17.2$ (b) $\mu(\gamma) \approx 37.5$ (c) $\|z\|_\infty^2 \approx 7.4 \times 10^{-7}$ (d) $\|z\|_\infty^2 \approx 4.5 \times 10^{-6}$

Fig. 1. Kernel function value estimation error for (a, b) $k(x, x') = \exp(\frac{-\|x-x'\|_2^2}{4})$ and (c, d) $k(x, x') = (\langle x, x' \rangle + 1)^3$ with $n = 10000$.



(a) $k(x, x') = \exp(\frac{-\|x-x'\|_2^2}{8})$ (b) $k(x, x') = (\langle x, x' \rangle + 2)^3$

Fig. 2. ROC curves for the USPS dataset.

detection while sacrificing an increase in false alarm rate. On the other hand, if we aim to achieve a low false alarm probability and $P_{FA} = \mathbb{P}(T(x) > \eta_{FA}|\mathcal{H}_0) = \tau_{FA}$ for the complete data, then we can use $T_{GauL}(x)$ or $T_{polyL}(x)$ in lieu of $T(x)$ to maintain the desired false alarm rate, which will in turn lead to the degradation of the probability of detection.

5. SIMULATION RESULTS

We first investigate the performance of the Gaussian (Fig. 1(a),1(b)) and polynomial (Fig. 1(c),1(d)) kernel value estimation. Recall that Theorem 1 and 2 state that the larger $\mu(x - x')$ or $\|x \circ x'\|_\infty$ is, the larger number of observations is needed to obtain reliable estimation. Therefore we consider two cases when $\mu(x - x')$ or $\|x \circ x'\|_\infty$ is either large or small in their corresponding kernels. We randomly generate 20 pairs of unit ℓ_2-norm vectors $x, x' \in \mathbb{R}^{10000}$ for each case. For each sample size $m \in [11, 500]$ and every pair, we sample 100 different Ω without replacement. The plots show the minimum, maximum and mean value of $g(x_\Omega, x'_\Omega) - k(x, x')$ over these 100 trials of all 20 pairs. The results shown in Fig. 1 bear out the analysis since the estimation errors decay rapidly as m increases. Moreover, when the value of $\mu(x - x')$ and $\|x \circ x'\|_\infty$ gets larger, we often see the estimation error growing for the same m.

Finally, we apply our approach for kernel matched subspace detection on USPS dataset which contains a collection of $n = 256$-dimensional handwritten digits. We randomly select 300 images (without replacement) from digit "6" as training samples to learn the subspace $\mathcal{S}$ and choose 800 images of digit "6" and "8" (400 samples each) for testing purposes. All these samples are normalized to unit ℓ_2 norms, then we add white Gaussian noise with different expected noise power ($\mathbb{E}[\|w\|_2^2] = n\sigma^2$) to the test samples. In our experiment, the noise power is set to be 0, 0.3 and 0.6 and the number of observations of test data is only 40% of the signal dimension. We compare the detection performance in the missing data setting with complete data case using two different kernels: Gaussian kernel (Fig. 2(a)) and polynomial kernel (Fig. 2(b)). The subspace dimension r is set to be 6 and 10 respectively. The receiver operating characteristics (ROC) curves are plotted in Fig. 2 to show quantitative performance. In both experiments, the curves for the missing test data are close to the ones for the complete test data with the same noise level. It can be also inferred that as the noise level increases, the difference between the areas under the curves (AUC) for complete and missing data grows.

6. CONCLUSION

We have presented a nonlinear version of the matched subspace detector in the presence of missing data in the input space. The detection results based on real data confirm the effectiveness and robustness of our method. The results in this paper can be extended to the problem of kernel subspace assignment of incomplete vectors.

7. REFERENCES

[1] H. Kwon and N. M. Nasrabadi, "Kernel matched subspace detectors for hyperspectral target detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 178–194, 2006.

[2] M. Rangaswamy, F. Lin, and K. Gerlach, "Robust adaptive signal processing methods for heterogeneous radar clutter scenarios," *Signal Processing*, vol. 84, no. 9, pp. 1653–1665, 2004.

[3] Z. Wang, G. R. Arce, and B. M. Sadler, "Subspace compressive detection for sparse signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process. (ICASSP)*, 2008, pp. 3873–3876.

[4] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Sig. Process.*, vol. 42, no. 8, pp. 2146–2157, 1994.

[5] F. Vincent, O. Besson, and C. Richard, "Matched subspace detection with hypothesis dependent noise power," *IEEE Trans. Sig. Process.*, vol. 56, no. 11, pp. 5713–5718, 2008.

[6] B. Schölkopf, A. J. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[7] L. Balzano, B. Recht, and R. Nowak, "High-dimensional matched subspace detection when data are missing," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, 2010, pp. 1638–1642.

[8] B. Eriksson, L. Balzano, and R. Nowak, "High-rank matrix completion," in *Proc. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 373–381.

[9] C. McDiarmid, "On the method of bounded differences," *Surveys in Combinatorics*, vol. 141, pp. 148–188, 1989.