TARGET DETECTION AND IDENTIFICATION USING CANONICAL CORRELATION ANALYSIS AND SUBSPACE PARTITIONING

Wei Wang and Tülay Adalı

University of Maryland Baltimore County
Department of CSEE
1000 Hilltop Circle, Baltimore, MD 21250

Darren Emge

US Army
Edgewood Chemical and Biological Center
Aberdeen Proving Grounds, MD 21010

ABSTRACT

We present a data-driven approach for target detection and identification based on a linear mixture model. Our aim is to determine the existence of certain targets in a mixture without specific information on the targets or the background, and to identify the targets from a given library. We use the maximum canonical correlation between the target set and the observations as the detection score, and use coefficients of the canonical vector to identify the indices of the present components from the given target library. The performance of the detector is enhanced using subspace partitioning on the target library. Both simulation and experimental results are presented to demonstrate the effectiveness of the proposed method in Raman spectroscopy for detection of surface-deposited chemical agents.

***Index Terms*—** target detection, identification, canonical correlation, subspace partitioning, Raman spectroscopy

1. INTRODUCTION

The aim of target detection in a linear mixture model is to determine if certain components exist in a given set of observations. Linear mixture model has been widely used in signal processing applications. It can be represented as

$$\mathbf{X} = \mathbf{S}\mathbf{A} + \mathbf{V},$$

where $\mathbf{X}$ is an observation matrix, $\mathbf{A}$ a matrix of mixing coefficients, and $\mathbf{S}$ the component matrix, and $\mathbf{V}$ a noise matrix. Here $\mathbf{X}$, $\mathbf{S}$, and $\mathbf{V} \in \mathbb{R}^{N\times M}$, and $\mathbf{A} \in \mathbb{R}^{M\times M}$ where M is the number of observations, which we assume is equal to the dimension of the signal subspace, and N is the length of observations.

The target in detection can be a specific component. Approaches such as generalized likelihood ratio test (GLRT) [1], and detection with correlation bound (DCB) [2] have been adopted and show satisfactory performance in these cases. What we address in this paper is a more challenging problem, in which the target is a library of components of interest, *i.e.*, the present components are from a given library without the knowledge of specific index information. This problem can be divided into two steps:

1. Hypothesis testing:
   Given $\mathbf{X} = \mathbf{S}\mathbf{A}+\mathbf{V}$ and spectrum library $\mathcal{T} = \{\mathbf{t}_1, \cdots, \mathbf{t}_L\}$, where L is the number of target components of interest, determine whether one or more components in $\mathcal{T}$ exist in the mixing components, *i.e.*,

   $$\mathcal{H}_0 : \mathcal{S} \cap \mathcal{T} = \phi$$
   $$\mathcal{H}_1 : \mathcal{S} \cap \mathcal{T} \neq \phi$$

   where ϕ denotes the empty set.

2. Identification:
   If $\mathcal{H}_1$, identify the index of the component that is present from the given library.

When background is known, and given that the library is ensured to cover all possible present components, supervised approaches such as GLRT or linear regression methods can be used [1] for the problem. In practice, however, it is usually difficult to obtain a reliable prior estimate of the background components, and the accuracy and comprehensiveness of the component library cannot be guaranteed, hence limiting the utility of these methods.

The aim of this paper is to develop a data-driven detection method without having to use *a priori* information. In practice, least squares (LS) or non-negative least squares (NNLS) methods have been used in applications such as Raman spectroscopy [1] where the interference of background is ignored in the detection. In this paper, we use the maximum canonical correlation between the target library and a block of the mixtures as the detection score, and use the coefficients of canonical vector to determine which components are present in the mixtures. Hence both the detection and the identification problems can be solved by the approach at the same time using detection with canonical correlation analysis (DCC).

In DCC, we use the target library as a projection subspace, hence its condition number is crucial to the performance of the detection algorithm. High canonical correlations between linear combinations of spectra are major causes for false positives as well as incorrect identifications of components that are actually present.

In this paper, we incorporate DCC detector with a library partitioning scheme by posing the problem as a vertex coloring problem in graph theory, in which linear combinations that cause high canonical correlations are regarded as adjacent vertices with different colors, hence leading to improvement of the performance by performing DCC on the partitioned subspace (DCC-P).

THIS PROJECT IS SUPPORTED BY EDGEWOOD CHEMICAL BIOLOGICAL CENTER, US ARMY RDECOM UNDER CONTRACT NO: W91ZLK-04-P-0950L

1-4244-1484-9/08/$25.00 ©2008 IEEE

We apply the proposed detection methods to Raman spectroscopy. A Raman spectrum gives a set of peaks that correspond to the characteristic vibrational frequencies of the material, which can be used as a signature for identification of various materials. We consider the application of Raman spectroscopy on non-contact detections of surface-deposited chemical agents, which is particularly useful for detecting environmentally hazardous chemicals [1]. Both the simulation and experimental results in Raman spectroscopy demonstrate the effectiveness of the proposed approach for the problem.

2. DETECTION USING CANONICAL CORRELATION

We investigate the relationship between the observation data set and the target library using canonical correlation analysis since it provides information on the closeness of two sets of vectors.

Given two sets of vectors, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_M] \in \mathbb{R}^{N\times M}$, and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_L] \in \mathbb{R}^{N\times L}$, canonical correlation analysis seeks a pair of vectors, $\mathbf{a}^*$ and $\mathbf{b}^*$, that maximize the correlation $\rho = \mathrm{corr}(\mathbf{Xa}, \mathbf{Yb})$, such that

$$\rho^* = \max_{\mathbf{a},\mathbf{b}} \mathrm{corr}(\mathbf{Xa}, \mathbf{Yb}). \quad (1)$$

The solution of Eq. (1) can be obtained by solving the following eigenvalue problems:

$$\begin{aligned} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{a}^* &= \rho^2\mathbf{a}^* \\ \mathbf{b}^* &= \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{a}^*, \end{aligned} \quad (2)$$

where $\mathbf{C}_{xx} = E\left[\mathbf{XX}^T\right]$, $\mathbf{C}_{yy} = E\left[\mathbf{YY}^T\right]$, $\mathbf{C}_{xy} = E\left[\mathbf{XY}^T\right]$, and $\mathbf{C}_{yx} = E\left[\mathbf{YX}^T\right]$.

The square roots of the eigenvalues obtained from Eq. (2) are called *canonical correlations*, and the vectors $\mathbf{a}^*$ and $\mathbf{b}^*$ *canonical vectors*.

To help explain the idea of DCC, we use a noiseless model $\mathbf{X} = \mathbf{SA}$, and let $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_L]$ be the target set. The maximum canonical correlation between $\mathbf{X}$ and $\mathbf{T}$ is given by

$$\rho^* = \max_{\mathbf{a},\mathbf{b}} \mathrm{corr}(\mathbf{Xa}, \mathbf{Tb}) = \max_{\mathbf{a},\mathbf{b}} \mathrm{corr}(\mathbf{SAa}, \mathbf{Tb}).$$

We can see that

- Under $\mathcal{H}_0$:
  $\mathcal{S} \cap \mathcal{T} = \phi$, hence
  $$\rho^* = 0$$
  if the subspaces spanned by $\mathbf{S}$ and $\mathbf{T}$ are orthogonal. Note that this orthogonality condition is just a simplification to emphasize the general idea for this example, and is not a requirement of the DCC method.

- Under $\mathcal{H}_1$:
  $\mathcal{S} \cap \mathcal{T} \neq \phi$. Let $\mathbf{s}_1 = \mathbf{t}_j$, *i.e.*, $\mathbf{S} = [\mathbf{t}_j, \mathbf{s}_2, \cdots, \mathbf{s}_M]$, then
  $$\rho^* = 1.$$
  An example solution is $\mathbf{a}^* = \mathbf{A}^{-1}[1, 0, \cdots, 0]^T$, and $\mathbf{b}^* = [0, \cdots, 1_{(j)}, \cdots, 0]^T$ since $\mathbf{SAa}^* = \mathbf{Tb}^* = \mathbf{t}_j$. Note that the non-zero element in $\mathbf{b}$ indicates the index of component that is in the mixture $\mathbf{X}$.

This observation suggests that we can use the maximum canonical correlation,

$$\rho^* = \max_{\mathbf{a},\mathbf{b}} \mathrm{corr}(\mathbf{Xa}, \mathbf{Tb}) \quad (3)$$

to solve the D-Set problem using the following two steps:

1. Use ρ^* for the hypothesis test to determine if any component in the library is present,
2. Use canonical vector, $\mathbf{b}^*$, to determine the indices of those that are present.

3. LIBRARY PARTITIONING WITH GRAPH COLORING

As seen in equations given in (2), the of spectrum library matrix $\mathbf{T}$ is used in the solutions of DCC methods as a projection subspace, hence its condition plays an important role on detection performance. A canonical correlation value close to one implies that a component in the library is approximately equal to a linear combination of other components, as a result, false positives and incorrect identifications might occur in detection. The following are two examples based on the spectrum library that we use in our Raman spectroscopy study, in which there are a total of 62 spectra, $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_{62}]$, where the first 50 are spectra of target chemicals of interest, and the last 12 are spectra of background materials.

- Example 1: *False positive*
  The canonical correlation value between $\mathbf{t}_{54}$ and $[\mathbf{t}_{27}, \mathbf{t}_{28}]$ is close to 1, *i.e.*, $\mathbf{t}_{54} \approx \alpha\mathbf{t}_{27} + \beta\mathbf{t}_{28}$, where α and β are scalars. Hence when $\mathbf{t}_{54}$ is background, high detection index is obtained if using the whole spectrum library because of the existence of the targets, $\mathbf{t}_{27}$ and $\mathbf{t}_{28}$ in the library.

- Example 2: *Misidentification*
  Similar to the example above, we have $\mathbf{t}_1 \approx \alpha\mathbf{t}_7 + \beta\mathbf{t}_6$. Hence when $\mathbf{t}_6$ and $\mathbf{t}_7$ are the mixing chemicals, $\mathbf{t}_1$ might be detected as the present chemical.

High canonical correlations also lead to an ill-conditioned component matrix, which is well known as numerically unstable and suffers from sensitivity to round-off errors in the computation.

Thus our objection to partition the library by splitting chemicals whose linear combinations cause high canonical correlations by putting them into different clusters. Most clustering algorithms are based on point-to-point distance measures, however, the canonical correlation is a measure on a point-to-set basis. Therefore clustering algorithms are not useful for our library partitioning problem.

To reduce canonical correlations among the spectrum library, we first take a close look at, for example, the component $\mathbf{t}_1$. The canonical correlation between $\mathbf{t}_1$ and $\mathbf{T}\backslash\mathbf{t}_1$ is $\rho_1^* = 0.9983$, and the mixing vector

$$\mathbf{b}_1 = [\cdots,\ 0.01,\ 0.06,\ 0.43_{(6)},\ 1.00_{(7)},\ 0.03,\ \cdots].$$

The subscripts denote the indices of the coefficients in the mixing vector, and $\mathbf{b}_1$ is normalized such that the maximum

Cluster #	Spectra in the each cluster						
1	4	33	36	38	42	54	60
2	6	8	20	21	34	52	53
3	3	11	15	23	35	48	49
4	12	19	24	28	40	50	55
5	2	5	18	25	41	43	62
6	14	22	27	32	39	47	59
7	1	9	10	17	44	46	57

Table 1. Spectrum distribution in clusters after partitioning

coefficient is equal to one. We can see that targets $\mathbf{t}_6$ and $\mathbf{t}_7$ contribute most significantly for the high value ρ_1^* attains. We call $\{\mathbf{t}_6, \mathbf{t}_7\}$ a *forbidden pair* of $\mathbf{t}_1$ since they together cause a high canonical correlation with $\mathbf{t}_1$. This also suggests that ρ_1^* can be decreased by breaking up the pair of $\mathbf{t}_6$ and $\mathbf{t}_7$, *i.e.,* putting $\mathbf{t}_6$ and $\mathbf{t}_7$ into different clusters. We continue finding such forbidden pairs for $\mathbf{t}_1$ until ρ_1^* is below a given threshold when all such pairs are split up.

After we find all forbidden pairs for each spectrum in the library, the next step is the partitioning of the library into a number of clusters such that the two elements of any forbidden pair are assigned into different clusters. This can be converted into a vertex coloring problem in graph theory.

In vertex coloring problem, different colors are assigned to the vertices of the graph such that no two *adjacent* vertices are assigned the same color. In graph theory, *adjacent* refers to vertices sharing the same edge. In our case, we consider those forbidden pairs as *adjacent* vertices, and want the number of colors to be as small as possible while satisfying the given constraints.

Graph coloring for an arbitrary graph is an NP-hard problem and has been well studied. A number of approximation and exact algorithms have been proposed. In our problem, library partitioning is a one-time procedure as long as the library does not change, hence an exact coloring algorithm is desirable and affordable. We implement an implicit enumeration algorithm using backtracking method [3].

The results of library partitioning are shown in Table 1, where each row corresponds to a cluster with its elements. Note that the goal of library partitioning is to reduce the canonical correlation with linear combinations of spectra. The single correlation (spectrum-to-spectrum) values are fixed and can not be decreased by any means. In our implementation, the spectrum pairs whose correlations are greater than the threshold are determined first, and one of the spectrum in each high correlated pair is extracted from the library before partitioning. In this report, 13 spectra are pulled out from the library. The information of extracted spectra is stored in a list of pairs. Whenever a spectrum in the list is detected, a second stage classification is performed to identify the present chemical between this spectrum and its counterpart.

To evaluate the condition of partitioned library, we need to calculate both the canonical correlations of each spectrum within its cluster and between the other clusters after partitioning.

The intra-cluster canonical correlation for spectrum $\mathbf{t}_i$ is defined as

$$\rho_i^{\text{intra}} = \max_{\mathbf{b}} \operatorname{corr}\left(\mathbf{t}_i, \left(\mathbf{T}^{(i)} \backslash \mathbf{t}_i\right) \mathbf{b}\right), \quad i = 1, \cdots, 62,$$

where $\mathbf{T}^{(i)}$ denotes the cluster to which $\mathbf{t}_i$ belongs.

The inter-cluster canonical correlation for $\mathbf{t}_i$ with cluster j is defined as:

$$\rho_{i,j}^{\text{inter}} = \max_{\mathbf{b}} \operatorname{corr}\left(\mathbf{t}_i, \mathbf{T}_j \mathbf{b}\right), \quad i = 1, \cdots, 62,\ j = 1, \cdots, M,$$

where $\mathbf{T}_j$ is the j-th cluster, $\mathbf{T} = \cup_{j=1}^{M} \mathbf{T}_j$, and M is the number of total clusters, and $i \notin \mathbf{T}_j$.

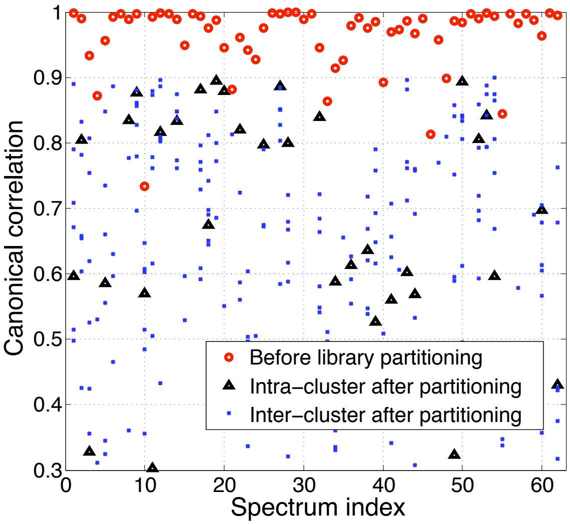

Fig. 1. Canonical correlations of each spectrum in the library before partitioning

Fig. 1 shows both inter- and intra-cluster canonical correlations for each spectrum in the library after partitioning, as well as intra-cluster canonical correlation before partitioning where the whole library is regarded as a cluster. Since there are a total seven clusters, there is one ρ_i^{intra} and six $\rho_{i,j}^{\text{inter}}$ values for each spectrum after partitioning in the library. We can see that a lot of canonical correlations are close to one before partitioning, and all canonical correlation values are decreased below the selected threshold of 0.9 after partitioning, thus decreasing the probability of false positives and misidentifications.

In DCC-P, DCC detector is performed on all clusters of the spectrum library, and the maximum DCC score is chosen as the DCC-P score.

4. SIMULATION AND EXPERIMENTAL RESULTS IN RAMAN SPECTROSCOPY

In simulations, we randomly create mixing matrices of which the coefficients follow a uniform distribution from [0, 1]. Noise is generated using Gaussian distribution. The signal-to-noise-ratio (SNR) is defined as: $\text{SNR} = 10\log_{10}(\frac{\|\mathbf{x}\|^2}{\sigma^2})$, where $\|\mathbf{x}\|^2$ is the average energy of the observation vector in $\mathbf{X}$, and σ^2 is the variance of the noise.

The detection performances of DCC, DCC-P, NNLS and NNLS-P are evaluated by the receiver operating characteristic (ROC) curves shown in Fig. 2. P_{FA} is the probability of false

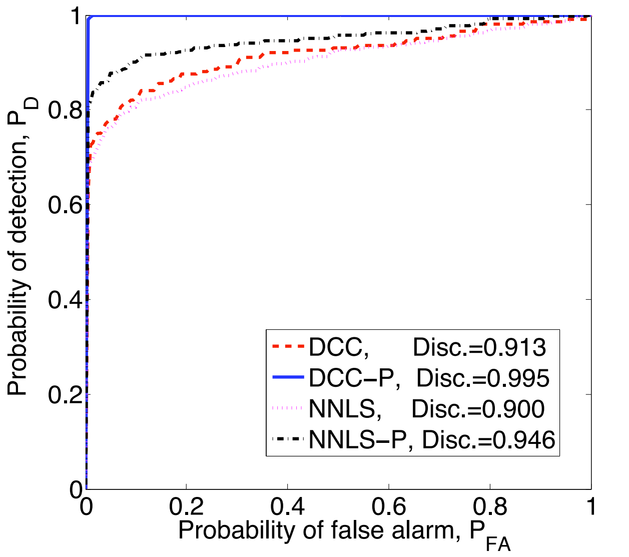

Fig. 2. ROCs for DCC, DCC-P, NNLS, and NNLSP (Present chemical=$\mathbf{t}_1$, background=$\mathbf{t}_{56}$, SNR= 5 dB)

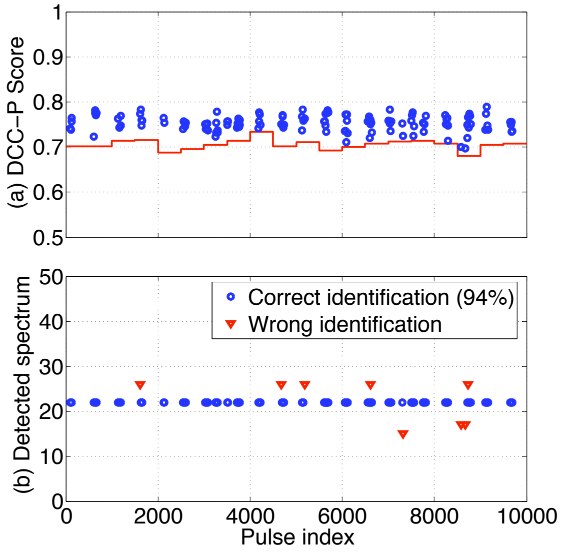

Fig. 3. Detection results of a laboratory experiment

alarm, or 1−specificity, and P_D is the probability of detection, or sensitivity. The area under the ROC curve measures *discrimination*, which is the ability of the test to make correct decisions. The discrimination values are given in each ROC plot.

In Fig. 2, we use $\mathbf{t}_{56}$ as the background, and $\mathbf{t}_1$ as the target chemical. The SNR is 5 dB. For each detection run, we use a block of 2 observations to form $\mathbf{X}$. Each curve is drawn using 200 runs.

We can see in Fig. 2 that DCC outperforms NNLS, and library partitioning improves detection performances of both the DCC and the NNLS detectors. The ROCs and discrimination values demonstrate the effectiveness of using the maximum canonical correlation as detection index in DCC. We also calculate the coefficients of the canonical vector $\mathbf{b}$ for each library spectrum in DCC-P, and first normalize the mean value of each element by its standard deviation, then divide the vector by its maximum value. The resulting largest elements in the canonical vector are given by

$$\mathbf{b} = [1.00, 0.34, \cdots, 0.09_{(9)}, \cdots, 0.09_{(17)}, \cdots],$$

where the subscript denotes the index of corresponding element in the 50-dimensional $\mathbf{b}$. The indices of the largest coefficient in $\mathbf{b}$ indicate that the present chemicals is $\mathbf{t}_1$, which is the correct identification in this simulation.

We also examine DCC-P with a total of 10000 observations in a laboratory experiment, where chemical MES (the 22-th spectrum in the library $\mathbf{T}$) is dropped one segment of a asphalt background. The solid line in Fig. 3(a) is a threshold calculated from estimated background samples for each block of 500 observations. A block-size of 10 observations is used for DCC-P detector. Since the present positions of MES are unknown, we are not able to obtain a ROC curve. The obvious periodic pattern in Fig. 3(a), however, implies a successful detection of DCC-P since the chemical is on a rotating platform, and the correct identification rate is satisfactory in Fig. 3(b).

5. CONCLUSION

In this paper, we propose a data-driven detection method for subset target detection based on a linear mixture model. By investigating canonical correlations and vectors between the mixtures and the target library, we can both detect and identify the present components from a given target library. The detector is incorporated with library partitioning to improve the detection performance. Additional improvement can be obtained by imposing a non-negativity constraint on CCA in applications such as Raman spectroscopy and image processing where contributions of mixing components can only be non-negative [4]. Both simulation and experimental results in Raman spectroscopy demonstrate the effectiveness of the proposed method.

6. REFERENCES

[1] ITT Industries, “Tests of the laser interrogation of surface agents system for on-the-move standoff sensing of chemical agents,” in *Proc. Int. Symp. Spectral Sensing Research*, 2003.

[2] W. Wang and T. Adalı, “Detection using correlation bound in a linear mixture model,” *Signal Processing,* vol. 87, no. 5, pp. 1118–1127, May 2007.

[3] M. Kubale and B. Jackowski, “A generalized implicit enumeration algorithm for graph coloring,” *Communications of the ACM*, vol. 28, pp. 412–418, 1985.

[4] W. Wang, T. Adalı, and D. Emge, “Unsupervised detection using canonical correlation analysis and its application to Raman spectroscopy,” in *Proc. IEEE Workshop on Machine Learning for Signal Processing*, Thessaloniki, Greece, Aug. 2007.