TARGET CONCEPT LEARNING FROM AMBIGUOUSLY LABELED DATA

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Changzhe Jiao

Dr. Alina Zare, Dissertation Supervisor

December 2017

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

TARGET CONCEPT LEARNING FROM AMBIGUOUSLY LABELED DATA

presented by Changzhe Jiao,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Alina Zare

Dr. James Keller

Dr. Marjorie Skubic

Dr. Dominic Ho

Dr. Ronald McGarvey

ACKNOWLEDGMENTS

I would like to express my special gratitude and thanks to my advisor, Dr. Alina Zare, for her invaluable guidance, support and encouragement throughout my studies and research. I would also like to thank my committee members, Dr. James Keller, Dr. Marjorie Skubic, Dr. Dominic Ho and Dr. Ronald McGarvey, for all of their insightful comments and suggestions.

I also want to thank my friends, former and current labmates, particularly to Shanjie Chen, Da Li, Xiaoxiao Du, Hao Sun, Piyush Khopkar and Matthew Cook, for our constructive discussions during my research work.

Thank you to my family for their endless love, support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF SYMBOLS

Symbol	Description
$\mathbf{x}$	Data vector/instance
N	Total number of data vectors/instances
n	Number of feature dimensions
$\mathbf{X}$	Data matrix, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$
$\mathbf{s}$	Target signature/endmember
$\mathbf{d}$	Estimated dictionary atom/endmember
$.^+$	Superscript for positive/target class
$.^-$	Superscript for negative/non-target class
T	Total number of estimated target dictionary atoms/endmembers
$.^T$	Superscript for Transpose
M	Total number of estimated non-target dictionary atoms/endmembers
$\mathbf{D}$	Dictionary/endmember matrix, $\mathbf{D} = [\mathbf{D}^+ \ \mathbf{D}^-] = [\mathbf{d}_1^+, \mathbf{d}_2^+, \cdots, \mathbf{d}_T^+, \mathbf{d}_1^-, \mathbf{d}_2^-, \cdots, \mathbf{d}_M^-]$
$\mathbf{a}$	Abundance vector or sparse codes of $\mathbf{x}$ corresponding to $\mathbf{D}$
$\mathbf{p}$	Abundance vector or sparse codes of $\mathbf{x}$ corresponding to $\mathbf{D}^-$
$\mathbf{A}$	Sparse codes/proportion matrix, $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N]$
$\boldsymbol{\mu}$	Data mean
$\boldsymbol{\Sigma}$	Data covariance matrix
$\mathcal{N}$	Normal distribution
λ	Sparsity level

$\mathbf{B}$	Data bag
K	Number of data bags
$\mathbf{x}_{ij}$	jth instance from the ith bag
L	Bag-level label
l	Instance-level label
$\mathbf{w}$	Weights for a linear classifier
C	Total number of classes
$\mathbf{W}$	Weights matrix for a multi-class linear classifier, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_C]$
δ	Step length for gradient descent
$\Lambda(\mathbf{x})$	Detector response of $\mathbf{x}$
$\mathbf{0}_{n\times 1}$	Column vector with number n 0 elements
$\mathbf{1}_{1\times M}$	Row vector with number M 1 elements
$\begin{bmatrix}\mathbf{A} & \mathbf{B}\end{bmatrix}$	Concatenating of arrays $\mathbf{A}$ and $\mathbf{B}$ horizontally
$\begin{bmatrix}\mathbf{A} \\ \mathbf{B}\end{bmatrix}$	Concatenating of arrays $\mathbf{A}$ and $\mathbf{B}$ vertically
$\mathbf{R}_N(\mathbf{a}) = [\mathbf{a}, \cdots, \mathbf{a}]_{1\times N}$	Matrix containing repeated entries of $\mathbf{a}$ with repetition of N vectors
$sign(\cdot)$	Sign function

LIST OF ACRONYMS

Acronym	Description
MIL	Multiple Instance Learning
eFUMI	Extended FUnctions of Multiple Instances
DL-FUMI	Dictionary Learning using FUnctions of Multiple Instances
SVM	Support Vector Machine
mi-SVM	mixed-integer SVM (instance-level)
MI-SVM	Mixed-Integer SVM (bag-level)
SMF	Spectral Matched Filter
ACE	Adaptive Coherence/Cosine Estimator
HD/HSD	Hybrid Detector/Hybrid Structured Detector
K-SVD	Dictionary Learning using Singular Value Decomposition
GPS	Global Positioning System
EM	Expectation Maximization
DD	Diverse Density
EM-DD	Diverse Density using Expectation Maximization
MI-HE	Multiple Instance Hybrid Estimator
GLRT	Generalized Likelihood Ratio Test
HBS	Hydraulic Bed Sensor
BCG	Ballistocardiogram
ECG	Electrocardiography
BMI	Body Mass Index
PD	Probability of Detection
FAR	False Alarm Rate

PFA	Probability of False Alarm
FCLS	Fully Constrained Least Squares
APR	Axis-Parallel Rectangles
DMIL/GDMIL	Dictionary based Multiple Instance Learning/ Generalized Dictionaries for Multiple Instance Learning
MILES	Multiple-Instance Learning via Embedded Instance Selection
VCA	Vertex Component Analysis
ISTA	Iterative Shrinkage Thresholding Algorithm
AUC/NAUC	Area Under Curve/ Normalized Area Under Curve
ROC	Receiver Operating Characteristic
NEON	National Ecological Observatory Network

ABSTRACT

The multiple instance learning problem addresses the case where training data comes with label ambiguity, *i.e.*, the learner has access only to inaccurately labeled data. For example, in target detection from remotely sensed hyperspectral imagery, targets are usually sub-pixel and the ground truthing of the targets according to GPS coordinates could drift across several meters. Thus the locations of the targets corresponding to the hyperspectral image are inaccurate. Training a supervised algorithm or extracting target signatures from this kind of labels is intractable. This dissertation investigates the topic target concept learning from ambiguously labeled data comprehensively; reviews and proposes several methods that either learn a set of representative or discriminative target concepts.

The multiple instance hybrid estimator (MI-HE) maximizes the response of the hybrid detector under a generalized mean framework and estimates a set of discriminative target concepts. MI-HE adopts a linear mixture model and iterates between estimating a set of discriminative target and non-target signatures and solving a sparse unmixing problem. MI-HE preserves bag-level label information for each positive bag and is able to estimate a target concept that is commonly shared among positive bags. Furthermore, MI-HE has the potential to learn multiple signatures to address signature variability.

After learning target concept, signature based detector could be applied for target detection. The presented algorithms were tested in many applications including simulated and real hyperspectral target detection, heartbeat characterization from ballistocardiogram signals and tree species classification from remotely sensed data. The presented algorithms were proven to be effective in learning high-quality target signatures and consistently achieved superior performance over the state-of-the-art comparison algorithms.

Chapter 1

Introduction

In supervised learning, each training data is assumed to be coupled with the desired classification output. However, acquiring accurately labeled training data can be time consuming, expensive and even infeasible. Furthermore, labeling ambiguity comes naturally in many machine learning and computer vision applications, for example, an image that is labeled as computer may also contain a desk or several books; a video that is labeled as abnormal may only have its subset frames containing an accident, making the training label ambiguous [2, 3].

In hyperspectral target detection [4, 5], ground truth label information coming from a GPS receiver could drift across several pixels depending on the accuracy of the GPS, thus it is only known that some area denoted by the ground truth contains some points of interest for sure. In medical applications like heartbeat characterization and heart rate estimation from Ballistocardiogram (BCG) signals [6–8], ground truth is not strictly aligned in time with the BCG signals and moreover, there may be some missed collection of heartbeat signals by the BCG sensors. These labeling uncertainties make traditional supervised learning

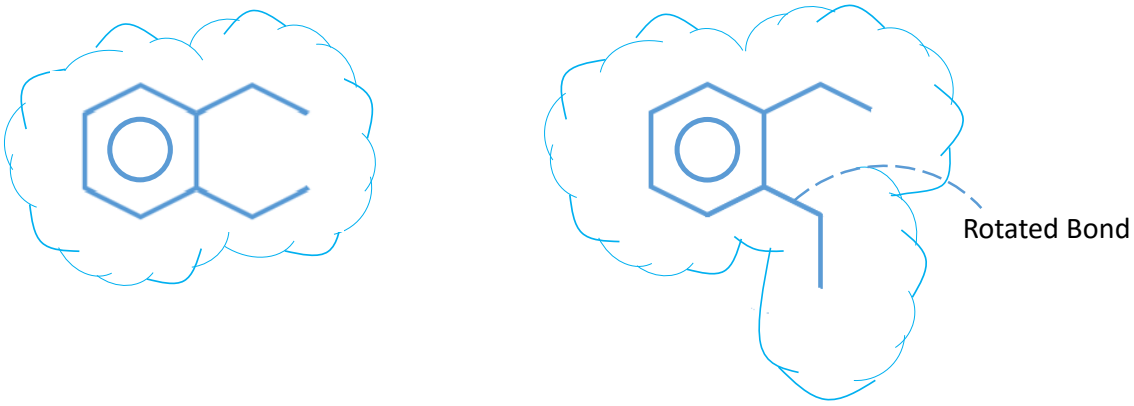


Figure 1.1: Illustration of MIL: a molecule with different shapes [1]

algorithms challenging to apply and the multiple instance learning algorithms more appealing.

Multiple instance learning (MIL) problem was first comprehensively investigated by Dietterich *et al.* [1] in the 1990s for the prediction of drug activity (musk activity). The effectiveness of a type of drug is determined by how tightly the drug molecule binds to a much larger protein molecule (*eg.*, enzymes and cell-surface receptors). However, a certain molecule determined by laboratory assay to be effective can have alternative variants called "conformations" - different structures the molecule could be by rotating its bonds shown by Fig. 1.1. Among all those different conformations the effective molecule could adopt, only one (or a few) actually binds to the desired target binding site. The learning task is to infer the correct shape of that molecule that actually has tight binding capacity.

In order to solve this problem, Dietterich *et al.* introduced the concept of "bags". Each molecule was treated as a bag and each possible conformation the molecule could be was treated as an instance in that bag. This directly induces the definition of multiple instance learning problem: a positively labeled bag contains at least one positive instance and neg-

atively labeled bags are composed of entirely negative instances. Finally, Dietterich *et al.* proposed to solve this problem by finding axis-parallel rectangles constructed by the conjunction of the features as approximation of the binding conformation.

Dietterich *et al.* also compared the proposed algorithm with several classical supervised learning algorithm including the backpropagation neural network and decision tree, and concluded that any supervised machine learning algorithm will perform poorly on MIL problem without considering the essence of MIL. Since Dietterich's work, many MIL learning algorithms were proposed and investigated. The MIL algorithms in the literature can be generally divided into two categories: learning an individual or a set of concepts that describe the positive class or learning a classifier that is able to classify individual instances or bags. This dissertation focuses on the former category, learning an individual or a set of concepts that either try to describe the positive class or distinguish the positive instances. Here, concepts refer to generalized class prototypes in the feature space.

1.1 Hyperspectral Image Analysis

Hyperspectral imaging spectrometers (also referred to as hyperspectral sensors) collect electromagnetic energy scattered in the scene across hundreds or thousands of spectral bands, and thus capture both the spatial and spectral information [9]. The spectral information is a combination of the reflection and/or emission of sunlight across wavelength by objects on the ground, and contains the unique spectral characteristics of different materials [10, 11]. The wealth of spectral information in hyperspectral imagery enables the possibility to conduct sub-pixel analysis including target detection [12, 13], precision agriculture [14, 15], biomedical applications [16, 17] and others [11, 18, 19].

1.1.1 Hyperspectral Image Data

Hyperspectral cameras collect radiance data over a high resolution range of wavelength, typically in the range of 0.3 μm to 2.5 μm [20] and construct a three-dimensional data cube. In hyperspectral data cube, each layer corresponds to a certain band of wavelength over all pixels and each pixel corresponds to the radiance value at certain location over the entire spectral bands. Due to the spatial resolution of hyperspectral cameras and high diversity of nature scene, individual pixel may be a mixture of several objects, in other words, each pixel may contain several different materials, called *endmembers*. Endmembers are assumed spectral value vectors over the wavelength for the pure materials present in the image. But the definition of "pure materials" could be also task driven or user defined.

As each pixel is a mixture of endmembers, abundances or proportions, are the amount or percentage of each endmember presents in an individual pixel. The magnitude of each endmember's proportion in an individual pixel is determined by many factors, *e.g.*, the relative area of the corresponding object, reflective intensity of materials, interactive absorption and scattering of light. Beside this, how the mixture is modeled also matters. Both linear and non-linear mixture models have been developed and verified to be effective in different physical context in the literature [11]. In realistic, the spectral mixture in remote sensing should be non-linear, due to the multiple mixture of light among different objects on the ground, *e.g.*, between tree canopy and the ground, and microscopic scattering between molecules. However, the linear mixing model that assumes each pixel is a convex combination of endmembers and proportions maintains the advantages of simplicity and good generalization ability and is investigated and adopted immensely. This dissertation mainly focuses on the linear mixing model.

1.1.2 Hyperspectral Unmixing

Hyperspectral unmixing can be decomposed into two major tasks: *endmember estimation* and *abundance estimation*. A mixing model needs to be assumed before conducting spectral unmixing. The convex mixing model assumes each pixel is a convex combination of the endmembers,

$$\mathbf{x}_j = \sum_{k=1}^{M} a_{jk}\mathbf{d}_k + \boldsymbol{\varepsilon}_j, j = 1, \ldots, N \tag{1.1}$$

$$\sum_{k=1}^{M} a_{jk} = 1, a_{jk} \geq 0, \forall j, k, \tag{1.2}$$

where N is the total number of data points, M is the number of endmembers (or materials), $\mathbf{x}_j$ is the spectral value of the j^{th} data point, $\boldsymbol{\varepsilon}_j$ is an error/noise term, $\mathbf{d}_k$ is the spectral signature of the k^{th} endmember, and a_{jk} is the abundance of the j^{th} pixel corresponding to the k^{th} endmember. The abundances in this model are constrained to the sum-to-one and non-negative constraint shown in Eq. (1.2). Typically, only the N data points are known as the input hyperspectral image, the remaining variables in the model including the spectral value of the endmembers, the number of endmembers, M, and corresponding abundance values are unknown and need to be solved. Estimating these unknown variables is an ill-posed inverse problem.

Many unsupervised hyperspectral unmixing methods adopt a number of assumptions about hyperspectral imagery to solve the ill-posed problem [10, 21–25]. For example, these methods include requiring the solution of endmembers to be found within the input data [26–31], adding volume penalty [32–35], assuming sparsity constraints [36–41], or adding the spatial smooth constrain on the abundance values [42–46]. These hyperspectral unmixing methods are mainly unsupervised algorithms. However, it is more appealing to

apply supervised or task driven unmixing [47, 48] if prior information about the particular materials of interest is available.

1.1.3 Hyperspectral Target Detection

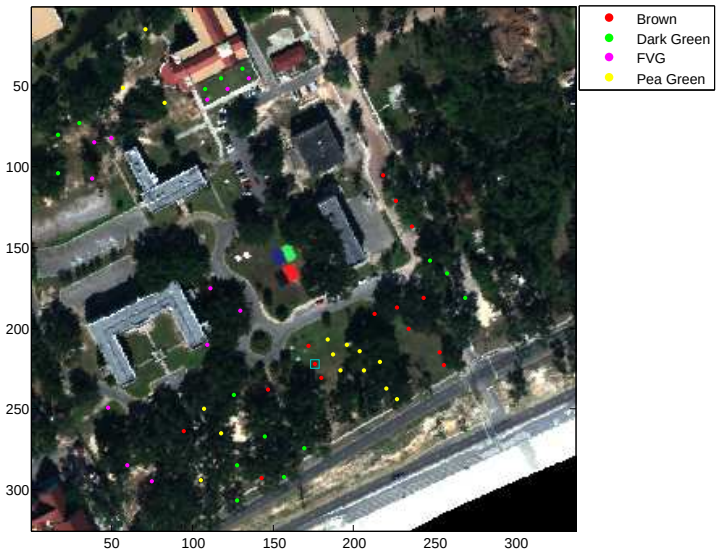
Hyperspectral target detection generally refers to the task of locating all instances of a target given a known spectral signature within a hyperspectral scene. A large number of hyperspectral target detection methods have been developed in the literature [4, 5, 49, 50]. The reasons most classification methods are not applicable to hyperspectral target detection tasks are threefold:

1. The number of training instances from the positive (target) class is small compared to that of the negative training data such that training an effective classifier is difficult. Typically in a hyperspectral image with size hundreds by hundreds pixels, there are only a few pixel or sup-pixel level target points. Compared with the number of the non-target points, this number of target points is too few to effectively train an classifier, *e.g.*, a SVM may be biased by the excessive non-target points and achieves very high classification accuracy but low detection rate.

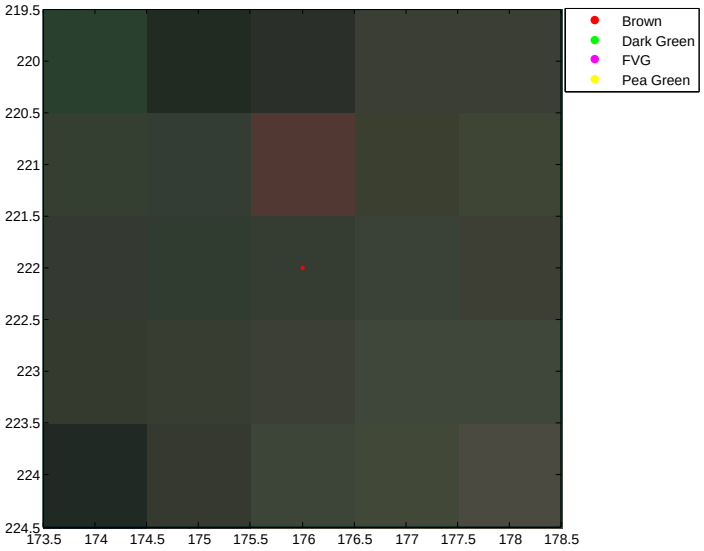
2. Due to the relatively low spatial resolution of hyperspectral imagery and the diversity of natural scenes, many targets are mixed points (sub-pixel targets). Most of the supervised learning algorithms assume each training data is a prime prototype of a class denoted by the label paired with this data. However, in hyperspectral image, a target pixel could be a mixture of several background materials and the amount of target mixture is unknown. Supervised learning algorithms will be stuck without considering the fact of mixture in training data.

3. Precise training labels are often difficult or infeasible to obtain. In hyperspectral image analysis, the ground truth information usually comes from a Global Positioning System (GPS) receiver placed to the target. However, the co-registration of the targets in the image to the GPS coordinates could drift for several meters. That means a target pixel denoted by the GPS coordinates could be a false positive point. The only reliable knowledge is with in a certain region there exists some targets for sure.

As an example, Fig. 1.2(a) shows the scattered target locations over MUUFL Gulfport data set collected over the University of Southern Mississippi-Gulfpark Campus [51], where there are 4 types of targets throughout the scene: Brown (15 examples), Dark Green (15 examples), Faux Vineyard Green (12 examples) and Pea Green (15 examples). The highlighted region shown in Fig. 1.2(a) is one of the brown target locations whose zoomed view is shown in Fig. 1.2(b), where we can clearly see that for this brown target there is one pixel drift between the real target location and ground truth location given by GPS. Developing a classifier or extracting a pure prototype for the target class given this incomplete knowledge of the training data is intractable, thus MIL methods are needed.



(a) Scattered target locations over MUUFL Gulfport data set



(b) Zoomed region of one target

Figure 1.2: Illustration of inaccurate coordinates from GPS: one target denoted as brown by GPS has one pixel drift.

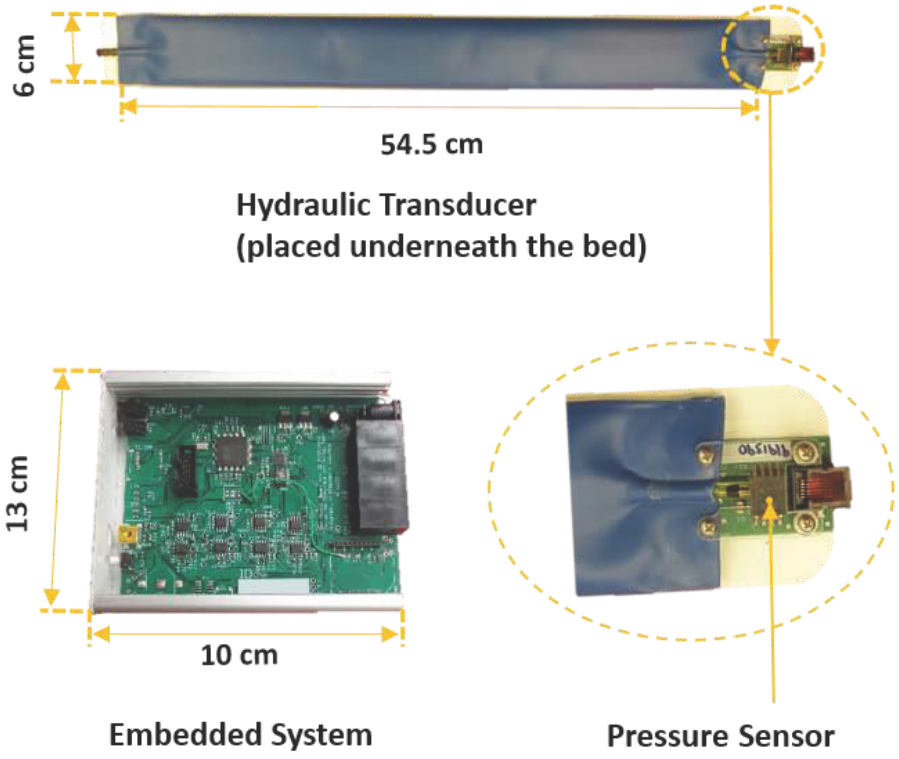
1.2 Ballistocardiogram Signal Analysis

Long-term measurement and monitoring of vital signs, *e.g.*, heart rate, respiratory rate, body temperature and blood pressure, provides promise for the early treatment of any potential problems, especially for older adults. Compared with the many wearable heart rate monitoring systems available, ballistocardiography provides an unobtrusive and, thus, comfortable monitoring alternative. These systems record the motion of the human body generated by the sudden ejection of blood into the large vessels at each cardiac cycle [6]. Such motion contains rich information and has gained revived interest due to recent development in measurement technology [7, 8] and a growing interest in managing chronic health conditions through passive sensors in the home [52].

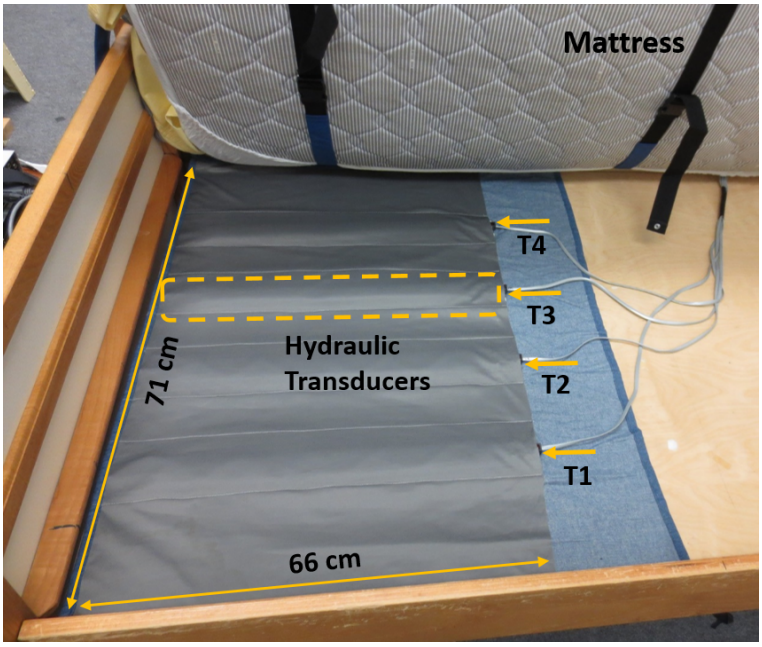
1.2.1 Hydraulic Bed Sensor System

The hydraulic bed sensor (HBS) developed at the Center for Eldercare and Rehabilitation Technology (CERT) at the University of Missouri is a BCG device providing a low-cost, noninvasive and robust solution for capturing physiological parameters during sleep [53–55]. The HBS was designed to maintain an imperceptible flat profile and to be used beneath a bed mattress. The system is comfortable for subjects lying on the mattress (*i.e.*, noninvasive), easy to install, watertight, and durable. Compared with other methods such as electrocardiography (ECG), BCG does not need electrodes or clips to be affixed to the patient's body and thus is ideal for long term in-home monitoring. However, the lack of saliency and large variability in a BCG signal makes it much more difficult to detect individual heartbeats than with an ECG.

The HBS is composed of a transducer and a pressure sensor as shown in Fig. 1.3(a). The



(a) Sensor and Embedded System



(b) Transducer placement

Figure 1.3: Hydraulic Bed Sensor System. (a) Hydraulic transducer (top) and embedded system (bottom). (b) Transducer placement

transducer was designed to be placed under the subject's upper torso. It is 54.5 cm long, 6 cm wide, and is filled with 0.4 liters of water [53–55]. The integrated silicon pressure sensor (Freescale MPX5010GP) attached to the end of the transducer is used for measuring the vibration of human body arising from each heartbeat. It captures the information of heartbeat together with respiration and motion artifact. The signal from each transducer is then amplified, filtered and sampled at 100 Hz. For ground-truthing, a piezoelectric pulse sensor (TN1012/ST, ADInstruments) attached to subject's finger was used to record the pulse ejected by a heartbeat.

In order to ensure enough coverage, four transducers are placed in parallel underneath a mattress as shown in Fig. 1.3(b). The four transducers are identical and independent, but the data quality collected by those four transducers could vary depending on the sleeping position, type of mattress (*e.g.*, material, thickness) and the physical characteristics of the subject (*e.g.*, age, body mass index (BMI)).

1.2.2 Multiple Instance Learning Problem in Ballistocardiograms

Fig. 1.4 shows a typical filtered BCG signal collected by one transducer and the corresponding finger sensor ground truth information, where the green circles denote every peak location of the filtered BCG signal. From Fig. 1.4, it can be seen that near the ground truth locations denoted by the finger sensor, there are prominent peak patterns measured by the BCG transducer corresponding to heartbeats. However, although all of the sensors are expected to be capturing each corresponding heartbeat signal simultaneously, there is unavoidable misalignment between the finger sensor and each of the BCG pressure sensors. Furthermore, depending on the location and position of the subject lying on the bed, which of these BCG sensors are able to capture a clear heartbeat signal is difficult to determine.

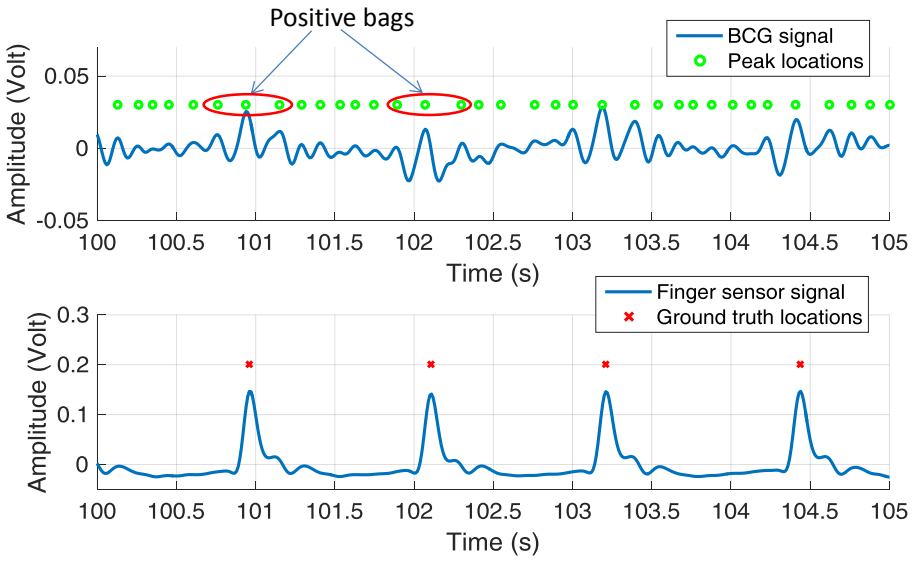


Figure 1.4: BCG signal and ground truth plot

These multiple labeling uncertainties in the training data cast more difficulties to traditional supervised learning methods for heartbeat detection and heart rate estimation from a BCG signal.

For this problem, this dissertation proposes to introduce the idea of training "bags" to address label uncertainty as well as mis-collection of heartbeat signals in the BCG data. So accurately labeled BCG signals are no longer needed. The proposed MI-HE algorithm is expected to learn a set of discriminative subject-specific heartbeat concepts from training bags of this type. After learning the heartbeat concept, a signature based detector can then be applied for real-time heartbeat monitoring and heart rate estimation.

1.3 Signature Based Detectors

The majority of sub-pixel detection techniques are statistical methods in which the target and background signals are modeled as random variables distributed according to some respective underlying probability distribution [4, 56, 57]. The detection problem can then be posed as a binary hypothesis test with two competing hypotheses: target absent ($\boldsymbol{H}_0$) or target present ($\boldsymbol{H}_1$) and a detector can be designed using the generalized likelihood ratio test (GLRT) approach [58]. Following the Neyman-Pearson criterion that maximizes the probability of detection (PD) given any desired probability of false alarm (PFA), the GLRT is shown in Eq. (1.3),

$$\Lambda(\mathbf{x}) = \frac{f(\mathbf{x}|\text{Target present})}{f(\mathbf{x}|\text{Target absent})} \triangleq \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{1.3}$$

where $f(\mathbf{x}|H_i)$ is the likelihood function value for each hypothesis.

1.3.1 Spectral Matched Filter

The hypotheses used for the spectral matched filter (SMF) [4, 58–61] are:

$$\begin{aligned} \mathbf{H}_0 &: \mathbf{x} \sim \mathcal{N}(0, \Sigma_b) \\ \mathbf{H}_1 &: \mathbf{x} \sim \mathcal{N}(a\mathbf{s}, \Sigma_b) \end{aligned} \tag{1.4}$$

where $\boldsymbol{\Sigma}_b$ is the background covariance and $\mathbf{s}$ is the known target signature which is scaled by a target abundance, a. The square-root of the GLRT for (1.4) results in the following as

the SMF detector:

$$\Lambda_{SMF}(\mathbf{x}, \mathbf{s}) = \frac{\mathbf{s}^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{x} - \boldsymbol{\mu}_b)}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{s}}} \tag{1.5}$$

where $\boldsymbol{\mu}_b$ is the background mean subtracted from the data to ensure a zero-mean background as defined in $\boldsymbol{H}_0$.

1.3.2 Adaptive Coherence/Cosine Estimator

The hypotheses used for the structured-background adaptive coherence/cosine estimator (ACE) [62–64] are:

$$\begin{aligned} \mathbf{H}_0 &: \mathbf{x} \sim \mathcal{N}\left(0, \sigma_0^2 \Sigma_b\right) \\ \mathbf{H}_1 &: \mathbf{x} \sim \mathcal{N}\left(a\mathbf{s}, \sigma_1^2 \Sigma_b\right) \end{aligned} \tag{1.6}$$

which includes $\sigma_0^2 = \frac{1}{n}\mathbf{x}^T \Sigma_b^{-1} \mathbf{x}$ and $\sigma_1^2 = \frac{1}{n}\left(\mathbf{x} - a\mathbf{s}\right)^T \Sigma_b^{-1}\left(\mathbf{x} - a\mathbf{s}\right)$ to add scale-invariance to the ACE detector where n is the dimensionality of the spectra. The square-root of the GLRT for (1.6) results in the following as the ACE detector [62, 63]:

$$\Lambda_{ACE}(\mathbf{x}, \mathbf{s}) = \frac{\mathbf{s}^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{x} - \boldsymbol{\mu}_b)}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{s}} \sqrt{(\mathbf{x} - \boldsymbol{\mu}_b)^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{x} - \boldsymbol{\mu}_b)}}. \tag{1.7}$$

Compared with Eq. (1.5), the ACE detector can be viewed as a normalized version of SMF: the input test points are whitened and normalized before the projection to the target signature. The normalization step removes the magnitude difference from the input data and achieves better performance in some scenarios, *e.g.*, test data with large variance in

magnitude.

1.3.3 Hybrid Detector

The hypotheses used for the background structured hybrid detector (HSD) [56, 65] are:

$$\begin{aligned} \mathbf{H}_0 &: \mathbf{x} \sim \mathcal{N}\left(\mathbf{D}^{-}\mathbf{p}, \sigma_0^2 \Sigma_b\right) \\ \mathbf{H}_1 &: \mathbf{x} \sim \mathcal{N}\left(\mathbf{D}\mathbf{a}, \sigma_1^2 \Sigma_b\right) \end{aligned} \tag{1.8}$$

where $\mathbf{D}$ and $\mathbf{D}^{-}$ represent the full endmember set and background endmembers set, respectively. $\mathbf{a}$ and $\mathbf{p}$ are the abundance values computed by Fully Constrained Least Squares (FCLS) [66] corresponding to $\mathbf{D}$ and $\mathbf{D}^{-}$, respectively. The GLRT for (1.8) results in HSD detector:

$$\Lambda_{HSD}(\mathbf{x}, \mathbf{D}) = \frac{(\mathbf{x} - \mathbf{D}^{-}\mathbf{p})^T \mathbf{\Sigma}_b^{-1} (\mathbf{x} - \mathbf{D}^{-}\mathbf{p})}{(\mathbf{x} - \mathbf{D}\mathbf{a})^T \mathbf{\Sigma}_b^{-1} (\mathbf{x} - \mathbf{D}\mathbf{a})}, \tag{1.9}$$

The hybrid detector models the reconstruction error of each point as a zero mean Gaussian distribution using the entire endmember set and non-target endmember set, respectively. The ratio between the reconstruction error using the entire endmember set and only the non-target endmember escalates the difference in the two reconstruction errors. The hybrid detector explicitly models the mixture in hyperspectral data and provides a sub-pixel detection alternative.

1.4 Overview of Research

In this dissertation, algorithms for target characterization (*i.e.*, estimation of target concept signatures) from training data with labeling ambiguity are presented. The goal of these algorithms are to estimate the target concept signatures from mixed training data that are effective for a follow-on target detection task. Since these algorithms extract the concept signatures from training data, then the background materials, environmental and atmospheric conditions, and other such variables are addressed during target characterization.

In the following, Chapter 2 provides a literature review of current multiple instance concept learning approaches and classifier learning approaches respectively. Chapter 3 introduces four previously proposed target concept learning algorithms: extended function of multiple instances (eFUMI) [67–70], dictionary learning using function of multiple instances (DL-FUMI) [71, 72], multiple instance spectral matched filter (MI-SMF) and multiple instance adaptive coherence/cosine estimator (MI-ACE) [73]. Chapter 4 investigates learning discriminative target concepts from MIL problem by maximizing the Hybrid Detector and proposes the multiple instance hybrid estimator (MI-HE) [74, 75]. Chapter 5 conducts a comprehensive testing of MI-HE and compares with previously proposed algorithms and the state-of-the-art MIL algorithms. Chapter 6 provides a conclusion and future work.

1.5 Formulation

Without loss of generality, let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{n \times N}$ be training data where n is the dimensionality of an instance and N is the total number of training instances. The data are grouped into K *bags*, $\mathbf{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$, with associated binary bag-level labels, $L =$

$\{L_1, \ldots, L_K\}$ where $L_i \in \{0, 1\}$; N_i is the number of instances in bag $\mathbf{B}_i$ and $\mathbf{x}_{ij} \in \mathbf{B}_i$ denotes the j^{th} instance in bag $\mathbf{B}_i$ with instance-level label $l_{ij} \in \{0, 1\}$. When identifying the label on a certain bag or instance is important, the N training data are assumed to be partitioned into K^+ positive bags with total number of instances N^+, and K^- negative bags with total number of instances N^-. Thus $N = N^+ + N^- = \sum_{i=1}^{K^+} N_i + \sum_{i=K^++1}^{K^++K^-} N_i$ where N_i is the number of instances for the ith bag. A positive bag will be indicated as $\mathbf{B}_i^+$ with associated bag level label $L_i = 1$ containing instances $\mathbf{x}_{ij}$ with instance-level labels l_{ij}, $s.t.$ $\sum_{j=1}^{N_i} l_{ij} \geq 1$. Similarly, $\mathbf{B}_i^-$ denotes a negative bag with bag level label $L_i = 0$ and instance-level labels $l_{ij} = 0$ (-1 for SVM based algorithms), $j = 1, \cdots, N_i$.

Chapter 2

Literature Review

This chapter provides a review of existing MIL algorithms discussed into two categories, multiple instance concept Learning and multiple instance classifier Learning, respectively.

2.1 Multiple Instance Concept Learning

Multiple instance concept learning refers to learning a description for the positive class given the bag-level labeled training data from MIL problem. Normally some prior knowledge is assumed in this step, *e.g.*, the estimated concept should be close to least one instance in each positively labeled bag and far away form every instance in the negatively labeled bags; the estimated concept must be a bad representation of all the negative instances. The estimated target concepts have the physical meaning to tell the unique features for the positive class and can be applied for further applications, *e.g.*, classification or regression.

2.1.1 Axis-parallel Rectangles

The Axis-parallel Rectangles (APR) [1] algorithms were proposed by Dietterich *et al.* for drug activity prediction in the 1990s. An axis-parallel rectangle can be viewed as an overlap or aggregation region of true positive instances in the feature space. In APR algorithms, a lower bond and upper bond are estimated for the scope of positive (active) class. Three APRs, GFS elim-count (greedy feature selection elimination count), GFS kde (greedy feature selection kernel density estimation) and iterated-discrim (iterated discrimination) algorithms were investigated and compared in [1].

GFS elim-count APR

The GFS elim-count APR refers to finding an APR in a greedy manner starting from the inclusion of all positive instances. This algorithm first finds the "all-positive APR" that exactly covers all of the positive instances. Fig. 2.1 shows the "all-positive APR" as a solid line bounding box of the instances, where the unfilled markers represent feature vectors of active instances and filled markers represent negative instances. As shown in the figure, the all-positive APR may contain several negative examples. The next step is to eliminate those negative instances and keep the positive instances as many as possible. A greedy shrinkage procedure was performed, which first excludes the "cheapest" negative instance by counting the minimum number of positive instances that needs to be removed from the APR for each negative instance. The greedy algorithm iteratively excludes the negative instance with the least cost (*i.e.*, the negative instance associates with the least positive instance to be removed) until all negative instances within the all-positive APR are eliminated. The dashed box in Fig. 2.1 indicates the final shrinkage APR by elim-count.

As stated in [1], feature selection is necessary as the features for this application are ex-

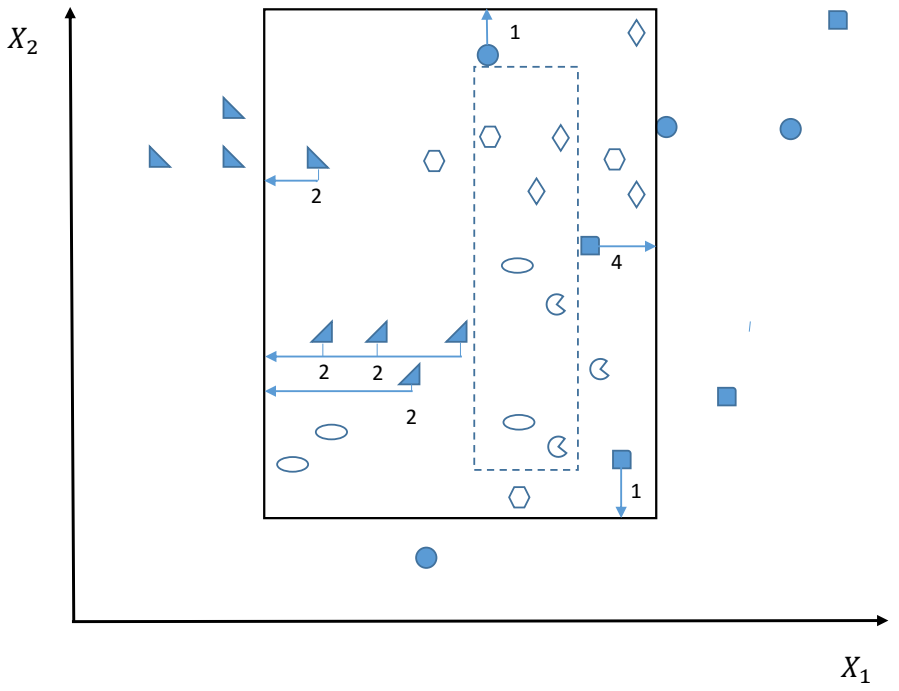


Figure 2.1: The elim-count procedure for excluding negative instances [1]

tracted by measuring the length of rays emanating from the origin of each instance (musk molecule) and nearby rays could be highly correlated. Also it is possible that only a subset of the feature dimensions is discriminative. So after constructing this shrinkage APR, the greedy feature selection algorithm that iteratively selects the feature dimensions that eliminates the most negative instances is conducted until no negative instance remain to be eliminated.

The GFS elim-count APR eliminates all negative instances from itself but one problem with this method is it is not guaranteed to contain at least one positive instance for each positive bag.

GFS kde APR

In order to solve the problem that GFS elim-count APR cannot preserve at least one positive instance for each positive bag, the author proposed to introduce a Gaussian kernel density estimate (kde) function to assign a cost value to each positive instance associated with the negative instance to be removed, GFS kde APR, instead of merely counting the number of positive instances must be eliminated for removal of one negative instance.

The proposed cost function is shown in Eq. (2.1), where $G_d(\mathbf{x}_{ij})$ is the Gaussian kernel density estimation denoting the probability of observing $\mathbf{x}_{ij}$. The cost function (2.1) adds three criteria to a positive instance associated with a negative instance to be removed:

1. The cost of removing $\mathbf{x}_{ij}$ should be small if there are many other positive instances $\mathbf{x}_{ik}, k = 1, \cdots N_i, k \neq j$, surviving in bag $\mathbf{B}_i^+$.

2. $\mathbf{x}_{ij}$ should be eliminated if there are many other positive instances are also observed at $\mathbf{x}_{ij}$, *e.g.*, the feature density at $\mathbf{x}_{ij}$ is high.

3. Assign a low cost value if $\mathbf{x}_{ij}$ is very isolated, *i.e.*, there are few other positive instances located near $\mathbf{x}_{ij}$ in the feature space.

$$\left(- \sum_{k=1, k \neq j}^{N_i} G_d(\mathbf{x}_{ij}) \right) + \alpha G_d(\mathbf{x}_{ij}) \tag{2.1}$$

According to Eq. (2.1), the last positive instance in each positive bag will be added by a very large cost value to not be eliminated. To some degree this "outside-in" method keeps the notion of MIL to have at least one positive instance per positive bag. However, one drawback of this algorithm is the computation complexity. It is quite expensive to compute every necessary kernel density estimates, *e.g.*, each negative instance may associate with several positive instances to be excluded across each of the n dimensions.

Iterated Discrimination APR

The iterated-discrim APR is an "inside-out" algorithm and tries to find the smallest APR that contains at least one instance per positive bag. It first choses an initial "seed" positive instance and iterates between two steps, growing a tight APR and selecting discriminating features until convergence, and then performs an expending procedure to improve its generalization ability, described as follows:

1. In this growing a tight APR step, the author proposed a cost function to define the size of an APR shown as Eq. (2.2), which is the sum of all its side length, where n is the index of feature dimension and ub_n and lb_n is the upper bond length and lower bond length of nth dimension, respectively. This cost function is optimized by a greedy algorithm to incorporate the "cheapest" positive instance followed by a

back-fitting algorithm [76] that tunes back at each greedy step.

$$Size(APR) = \sum_{n} ub_n - lb_n \tag{2.2}$$

2. In this feature selection step, the algorithm iteratively choose feature that "strongly discriminates" the most number of negative instances. Here the "strongly discriminate" is defined either if one negative instance lies more than 1 Å outside the bounds of the APR for feature dimension n or if one negative instance lies beyond the bounds of the APR and lies further along feature n than along any other dimensions.

   After iteration between step 1 and 2 (which is said to converge within 3-4 iterations), a too tight, sub-dimensional APR that excludes most positive instances was estimated. So a kernel density estimation method was adopted to expand this tight APR to include more positive instances which made the resulted APR a more generalized concept region. The iterated-discrim APR was verified to have the best performance on the musk dataset. However, one problem with the iterated-discrim APR is in theory the resulted APR may contain a subset of instances belong to negative bags.

2.1.2 Diversity Density

Diversity Density (DD) [2, 77] tries to learn a concept for the positive class that is close to the intersection of positive bags and far always from every negative instance, *i.e.*, an area preserves both high density of target points and low density of non-target points, called diversity density.

The proposed general maximum likelihood function by DD is shown in Eq. (2.3), where $\mathbf{s}$ is the assumed true concept for the positive class and $\mathbf{d}$ is the concept variable for

estimation,

$$\arg\max_{d} \prod_{i=1}^{K^+} \mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{B}_i^+) \prod_{i=K^++1}^{K^++K^-} \mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{B}_i^-) \tag{2.3}$$

Each term in the likelihood function Eq. (2.3) was defined by the noisy-or model,

$$\mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{B}_i^+) = \mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{x}_{i1},\mathbf{x}_{i2},\cdots,\mathbf{x}_{iN_i}) = 1-\prod_{j=1}^{N_i}(1-\mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{x}_{ij}\in\mathbf{B}_i^+)), \tag{2.4}$$

$$\mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{B}_i^-) = \prod_{j=1}^{N_i}(1-\mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{x}_{ij}\in\mathbf{B}_i^-)). \tag{2.5}$$

The casual probability for individual instance is modeled by the distance between the individual instance and the positive concept location,

$$\mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{x}_{ij}) = exp(-\|\mathbf{x}_{ij}-\mathbf{d}\|^2). \tag{2.6}$$

The intuitive understanding of the proposed noisy-or model is if there is at least one instance in positive bag $\mathbf{B}_i^+$ is close to $\mathbf{d}$ then $\mathbf{Pr}(\mathbf{d}=\mathbf{s}|\mathbf{B}_i^+)$ is high; thus the first term in the noisy-or model in Eq. (2.4) makes sure that the estimated $\mathbf{d}$ close to at least one instance in every $\mathbf{B}_i^+$. Eq. (2.5) drives the estimated $\mathbf{d}$ to be far away from every instance in $\mathbf{B}_i^-$. Similarly as stated in [1], a band selection was performed by optimization of the weights added to each dimension.

As stated by the author, the noisy-or model is highly non-smooth and there are several local maxima in the solution space which make finding the global optima very difficult. Gradient ascent with starting points from every positive instance was adopted to maximize the proposed log-likelihood function. Although it showed competitive performance to the comparison algorithms, the computational complexity is still a problem.

2.1.3 Expectation Maximization of Diversity Density

An expectation maximization version of Diversity Density (EM-DD) [78] was proposed by Zhang *et al.* in order to improve the computation time of DD [2, 77]. EM-DD assumes there exists only one instance per bag corresponding to the bag-level label and treats the knowledge of the key-point instance corresponding to the bag-level babel as a hidden latent variable. EM-DD starts with some initial guessing of the positive concept $\mathbf{d}$ and iterates between an expectation step (E-step) that picks one point per bag as the representative point of that bag and then performs a quasi-newton optimization (M-step) [79] on the single-instance DD problem. In more detail, in the E-step, the probability of each instance to be the one determines the bag-level label given target concept $\mathbf{d}$ from the previous iteration is estimated by a multivariate Gaussian distribution as shown in Eq. (2.7), where $\mathbf{x}_i^*$ is the assumed representative instance for bag $\mathbf{B}_i$. In the M-step, the positive concept $\mathbf{d}'$ is estimated by optimizing the standard DD problem with only one instance per bag determined in the E-step, shown in Eq. (2.8), where $\mathbf{Pr}(L_i|\mathbf{d}, \mathbf{x}_i^*)$ is the reduced single instance DD problem from Eq. (2.3).

$$\mathbf{x}_i^* = \arg \max_{\mathbf{x}_{ij} \in \mathbf{B}_i} exp(-\|\mathbf{x}_{ij} - \mathbf{d}\|^2) \tag{2.7}$$

$$\mathbf{d}' = \arg \max_{\mathbf{d}} \prod_i \mathbf{Pr}(L_i|\mathbf{d}, \mathbf{x}_i^*) \tag{2.8}$$

It was stated in the original paper "EM-DD runs over 10 times faster than DD on Musk 1 and over 100 times faster when applied to Musk2" [78] and achieved the highest accuracy (above 95%) over the comparison algorithms by picking specific initialization using validation data. However, EM-DD was later verified in [80] to have close but inferior

performance to DD.

2.1.4 Dictionary based Multiple Instance Learning

Dictionary based Multiple Instance Learning (DMIL) [81] and its generalization, Generalized Dictionaries for Multiple Instance Learning (GDMIL) [82], propose to optimize the noisy-or model using dictionary learning methods [83–87]. The target concept estimated is a set of dictionary atoms. In detail, the author models the probability of individual instance to be positive as a zero-mean multi-variate Gaussian distribution of the reconstruction error between the individual instance and the linear combination of positive dictionary atoms, shown as Eq. (2.9), where p_{ij} is the probability for instance $\mathbf{x}_{ij}$ to be a true positive point; $\mathbf{D}$ is the estimated dictionary set as the positive concept set and $\mathbf{a}_{ij}$ is the sparse representation of $\mathbf{x}_{ij}$ given $\mathbf{D}$,

$$p_{ij} \propto exp(-\|\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}\|_2^2). \tag{2.9}$$

Given the model defined in Eq. (2.9), the modified noisy-or model to be optimized is shown in Eq. (2.10) and the negative logarithm of Eq. (2.10) is shown in Eq. (2.11), where α is a scaling term to control the influence of negative bags.

$$J(\mathbf{D}, \mathbf{X}) = \prod_{i=1}^{K^+} \left(1 - \prod_{j=1}^{N_i}(1 - p_{ij})\right) \prod_{i=K^++1}^{K^++K^-} \left(\prod_{j=1}^{N_i}(1 - p_{ij})\right) \tag{2.10}$$

$$-\log J(\mathbf{D}, \mathbf{X}) = -\sum_{i=1}^{K^+} \log \left(1 - \prod_{j=1}^{N_i}(1 - p_{ij})\right) - \alpha \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} \log(1 - p_{ij}) \tag{2.11}$$

By substituting Eq. (2.9) to Eq. (2.11), the objective function is optimized iteratively

between two steps, a dictionary learning step that solves the dictionary $\mathbf{D}$ atom by atom using gradient descent method and a sparse coding step that solves the sparse representation of each instance in $\mathbf{X}$ given current dictionary $\mathbf{D}$ using orthogonal matching pursuit [88, 89].

The advantages of DMIL over the past multiple instance concept learning algorithms lie in two folds:

1. Instead of learning one concept for the positive class, DMIL learns a set of positive dictionary atoms to better describe the positive class.

2. The second term in the negative log-likelihood function, $-\alpha \sum_{i:L_i=-} \sum_{j=1}^{N_i} \log(1 - p_{ij})$, enforces that the negative instances are all poorly represented by the estimated dictionary $\mathbf{D}$, so that $\mathbf{D}$ maintains discriminative features of the positive class and contains the least information from the negative class.

2.2 Multiple Instance Classifier Learning

Multiple instance classifier learning refers to training a discriminative model from labeled training bags in MIL problem for prediction the label of unknown bags or individual instances. Since the positive bags are mixture of both positive and negative data, the multiple instance classifier learning algorithms in the literature typically train a classifier through a heuristic way, *i.e.*, starting from some initial guessing of the labels for data from positively labeled bags.

2.2.1 Mixed Integer Support Vector Machine

Andrews *et al.* model the MIL problem as a generalized mixed integer formulation of support vector machine [90] algorithms (mi-SVM and MI-SVM). The two proposed algorithms lead to mixed integer quadratic programming problem and were solved through heuristic ways. The two algorithms mi-SVM and MI-SVE differ in the manner of selection of training data. mi-SVM adopts the entire training data into consideration to train a SVM and modifies the instance-level label iteratively; whereas the MI-SVM trains a SVM by selecting one instance per bag with the maximum classification confidence as representative instance of each bag. The two algorithms stop when there is no change in the assigned label to instances across two iterations. mi-SVM and MI-SVM assume the labels for the training data are subject to the following MIL constraint shown in Eq. (2.12):

$$\sum_{\mathbf{x}_{ij} \in \mathbf{B}_i} \frac{l_{ij}+1}{2} \geq 1, \;\; \forall i \; s.t. \; L_i = 1, \text{ and } l_{ij} = -1, \; \forall i \; s.t. \; L_i = -1 \tag{2.12}$$

mi-SVM

The mi-SVM tries to solve a soft-margin maximization problem jointly over the possible labels assigned to individual instances and the hyperplane. The formulation of mi-SVM is shown in Eq. (2.13).

$$\begin{aligned} &\min_{\{l_{ij}\}} \min_{\mathbf{w},b,\xi_{ij}} \frac{1}{2}\|\mathbf{w}\|^2 + \alpha \sum_{i=1}^{N} \xi_{ij} \\ &s.t. \; \forall i : l_{ij}(\langle \mathbf{w}, \mathbf{x}_{ij} \rangle + b) \geq 1 - \xi_{ij}, \; \xi_{ij} \geq 0, \; l_{ij} \in \{-1, 1\}, \text{ and (2.12) holds,} \end{aligned} \tag{2.13}$$

where $(\mathbf{w}, b)$ are the weights and bias for a SVM classifier, ξ_{ij} is a slackness term and α is the scaling factor for slackness.

The problem of optimizing Eq. (2.13) is the accurate values of all l_{ij} from positive bags are not known. In order to solve this mixed-integer quadratic programming problem, the author adopted a heuristic optimization strategy. Specifically, the label l_{ij} for each instance from positive bags was initialized by generalizing the bag-level label L_i to individual instance, $l_{ij} = L_i$, for $L_i = 1$. Then a SVM was trained and applied to the positive training bags again to reset its instance-level label. If any of the positive bag has its all instances classified into negative, *i.e.*, $\sum_{\mathbf{x}_{ij} \in \mathbf{B}_i^+} (1 + l_{ij})/2 == 0$, the instance in this bag with maximum confidence value to be positive will be assigned a positive label and a SVM was trained again based on the newly reset labels. The algorithm stops until there is no change in the instance-level label.

The mi-SVM starts from all instances from positive bags with label 1 and iteratively modifies the labels according to the bag constraint in MIL problem until there is no change in label re-assignment. It finally looks for a MI-separating hyperplane such that each positive bag has at least one of its instances is classified into positive and all negative instances are separated to the other side of the hyperplane. However, there is no guarantee that this process will converge.

MI-SVM

Another way to applying max-margin optimization to mixed integer SVM is to generalize the notion of margin to be maximized from individual instances to bags. The functional margin for each bag, γ_I, is defined by the maximum decision values of the entire instances in each bag, shown as Eq. (2.14). In each iteration, only the instances with the maximum

decision value in each bag are adopted to train a SVM. One thing to notice is the margin of a positive bag is defined by the "most positive" instance, whereas the margin of a negative bag is defined by the "least negative" instance.

$$\gamma_i \equiv L_i \max_{\mathbf{x}_{ij} \in \mathbf{B}_i} (\langle \mathbf{w}, \mathbf{x}_{ij} \rangle + b) \tag{2.14}$$

The formulation of MI-SVM is shown in Eq. (2.15), where ξ_i is the slackness for bag $\mathbf{B}_i$. MI-SVM is initialized by assigning the mean of each bag as its representative training instance and optimized alternatively between training a SVM given the single training instance from each bag and picking a representative instance with the maximum decision value according to currently trained SVM. The algorithm stops until there is no change in the selection of training instance from each bag. However, similar to mi-SVM, the convergence of MI-SVM is also not guaranteed.

$$\begin{aligned} &\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + \alpha \sum_i \xi_i \\ &s.t.\ \forall i : L_i \max_{\mathbf{x}_{ij} \in \mathbf{B}_i} (\langle \mathbf{w}, \mathbf{x}_{ij} \rangle + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \end{aligned} \tag{2.15}$$

2.2.2 Multiple-Instance Learning via Embedded Instance Selection

The Multiple-Instance Learning via Embedded Instance Selection (MILES) [91] relaxes the constraint in MIL that negative bags are composed of all negative instances and allows target concept to be related to negative bags for a more general application in computer vision. For example, in objective recognition application, an instance (typically small patch in an image) is labeled as positive to indicate part of the object, however a negative bag may

also contain patches that look like parts of the object. MILES proposed to first embed each bag to a target concept based feature space, where the set of candidate target concept comes from the union of all bags. Then a 1-norm SVM [92] was trained on the feature vectors extracted from each bag; finally, an instance selection was performed based on the SVM decision value to realize instance-level classification.

MILES adopts each instance in the training bag as a candidate for target concepts, *i.e.*, $\mathcal{D} = \{\mathbf{x}_k : k = 1, \cdots, N\}$. The kth value of embedded feature vector for bag $\mathbf{B}_i$ corresponding to candidate target concept $\mathbf{x}_k$ is computed according to Eq. (2.16), where $s(\mathbf{x}_k, \mathbf{B}_i)$ is a similarity function.

$$\Pr(\mathbf{x}_k|\mathbf{B}_i) \propto s(\mathbf{x}_k, \mathbf{B}_i) = \max_{\mathbf{x}_{ij} \in \mathbf{B}_i} \exp(-\frac{\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2}{\sigma^2}) \tag{2.16}$$

By applying Eq. (2.16) to all candidate target concepts in $\mathcal{D}$, a bag $\mathbf{B}_i$ is then embedded into a N dimensional space $\mathbb{F}_{\mathcal{D}}$, with coordinate $\mathbf{v}(\mathbf{B}_i)$ shown in Eq. (2.17). Applying the mapping (2.17) embeds all training bags into $\mathbb{F}_{\mathcal{D}}$, as a $N \times (K^+ + K^-)$ matrix:

$$\mathbf{v}(\mathbf{B}_i) = \Big[s(\mathbf{x}_1, \mathbf{B}_i), s(\mathbf{x}_2, \mathbf{B}_i), \cdots, s(\mathbf{x}_N, \mathbf{B}_i) \Big]^T \tag{2.17}$$

After the mapping of training bags into $\mathbb{F}_{\mathcal{D}}$, a MIL problem was converted into a supervised leaning problem and was solved by 1 norm SVM [92] formulated as Eq. (2.18), where $(\mathbf{w}, b)$ are weights and bias for a linear SVM classifier and $\boldsymbol{\xi}$ are slackness. α_1 and α_2 are chosen differently to assign different penalty on false negatives and false positives.

$$
\begin{aligned}
\min_{\mathbf{w},b,\xi,\eta} \quad & \lambda \sum_{k=1}^{N} |w_k| + \alpha_1 \sum_{i=1}^{K^+} \xi_i + \alpha_2 \sum_{i=K^+ +1}^{K^+ + K^-} \xi_i \\
s.t. \quad & (\mathbf{w}^T \mathbf{v}_i^+ + b) + \xi_i \geq 1, i = 1, \cdots, K^+, \\
& -(\mathbf{w}^T \mathbf{v}_i^- + b) + \xi_i \geq 1, i = K^+ + 1, \cdots, K^+ + K^-, \\
& \xi_i \geq 0, i = 1, \cdots, K^+, K^+ + 1, \cdots, K^+ + K^-,
\end{aligned}
\tag{2.18}
$$

After solving the 1 norm SVM by linear programming (LP) [93, 94] for $(\mathbf{w}^*, b^*)$, the set of selected features, $\mathcal{I}$, was determined by the index set of nonzero values in $\mathbf{w}^*$, shown in Eq. (2.19). And the discriminant function for classifying a bag $\mathbf{B}_i$ is shown in Eq. (2.20). This completes the bag classification step.

$$
\mathcal{I} = \{k : |w_k^*| > 0\} \tag{2.19}
$$

$$
y = sign(\sum_{k \in \mathcal{I}} w_k^* s(\mathbf{x}_k, \mathbf{B}_i) + b^*) \tag{2.20}
$$

In some applications of MIL, instance-level classification is also required. For example, in objection detection, it is not enough to only identify if an image contains or not a target, telling where is the target is also crucial. After learning discriminant function for bags, MILES realizes instance-level classification by computing the contribution of individual instance to the classification of the bags. Specifically, an instance in a bag $\mathbf{B}_i$ has contribution to the discriminant function value $\sum_{k \in \mathcal{I}} w_k^* s(\mathbf{x}_k, \mathbf{B}_i) + b^*$ greater (or less) than an empirical threshold will be identified as positive (or negative).

2.2.3 Multi-Instance Dictionary Learning

The Multi-Instance Dictionary Learning (MIDL) [95] is a multiple instance dictionary learning algorithm for detection abnormal events in videos. The application context is in the public video surveillance application, it is difficult to label each video frame as normal (negative) or abnormal (positive), the only known labeling information is the segment of video that contains abnormal events. So each segment of video could be regarded as a bag and its bag level label is determined by if it contains abnormal events. Specifically, the author assumes there exists a set of dictionary $\mathbf{D} \in \mathbb{R}^{n \times M}$ that can better represent the training data in the sense of classification and the label of bag is determined by the instance in it with maximum classification value. The proposed objective is shown in Eq. (2.21),

$$\min_{\mathbf{D},\mathbf{W}} \frac{1}{K} \sum_{i=1}^{K} \log(1 + e^{-L_i \max_{j=1\cdots N_i} l(\mathbf{x}_{ij}, \mathbf{a}_{ij}, \mathbf{W})}) + \frac{\alpha}{2} \|\mathbf{W}\|_F^2, \tag{2.21}$$

where $\mathbf{a}_{ij}$ is the sparse representation of $\mathbf{x}_{ij}$ given $\mathbf{D}$, $l(\mathbf{x}_{ij}, \mathbf{a}_{ij}, \mathbf{W}) = \mathbf{x}_{ij}^T \mathbf{W} \mathbf{a}_{ij} + b$ is the discriminant function to classify each instance into positive or negative, $\mathbf{W} \in \mathbb{R}^{m \times k}$ is the classification weights matrix, $b \in \mathbb{R}$ is the bias and α is the scaling factor for weights $\mathbf{W} \in \mathbb{R}^{m \times k}$.

The variables are solved alternatively between the sparse codes $\mathbf{a}$, dictionary $\mathbf{D}$ and regression matrix $\mathbf{W}$. Specifically, $\mathbf{a}$ is solved as the least angle regression (LARS) problem [96] shown in Eq. (2.22) and $\mathbf{D}$ and $\mathbf{W}$ are solved by gradient descent by taking gradient on objective function (2.21). Note that the logistic regression $\log(1+e^{-L_i \max_{j=1\cdots N_i} l(\mathbf{x}_{ij}, \mathbf{a}_{ij}, \mathbf{W})})$ is convex but not smooth with respect to $\mathbf{W}$, so the sub-gradient was also used.

$$\mathbf{a}^*(\mathbf{x}, \mathbf{D}) \triangleq \arg\min_{\mathbf{a} \in \mathbb{R}^M} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\mathbf{a}\| + \lambda_1 \|\mathbf{a}\|_1 + \frac{\lambda_2}{2} \|\mathbf{a}\|_2^2 \tag{2.22}$$

2.2.4 Max-Margin Multiple-Instance Dictionary Learning

The Max-Margin Multiple-Instance Dictionary Learning (MMDL) [97] adopts the idea of bag of words (BoW) model [98] and trains a set of linear SVMs as codebook. The novel assumption of MMDL is the positive instances could belong to many different clusters. The motivation of this assumption lies in the truth in computer vision that positive class may have many different categories. For example, the positive class "computer room" may have image patches containing desk, screen, keyboard.

MMDL assumes there exists a latent variable for each instance denoting its cluster, $z_{ij} \in 0, 1, \cdots, C$, where C is the assumed number of positive classes. For each instance $\mathbf{x}_{ij}$, $z_{ij} = 0$ denotes this instance is from the negative class; otherwise, $\mathbf{x}_{ij}$ is from the cth positive class given $z_{ij} = c$, $c = 1, \cdots, C$. Furthermore, a set of linear SVM classifiers were also introduced as a matrix with each its column as a weight vector, $\mathbf{W} = [\mathbf{w}_0, \mathbf{w}_1 \cdots, \mathbf{w}_C]$, $\mathbf{w}_c \in \mathbb{R}^{n \times 1}$, $c \in \{0, 1, \cdots, K = C\}$. The cluster of instance $\mathbf{x}_{ij}$ is determined according to Eq. (2.23):

$$z_{ij} = \arg\max_c \mathbf{w}_c^T \mathbf{x}_{ij} \tag{2.23}$$

The proposed formulation of MMDL is shown in Eq. (2.24),

$$\begin{aligned} \min_{\mathbf{W}, z_{ij}} \quad & \sum_{c=0}^{C} \|w_c\|^2 + \alpha \sum_{i=1}^{K} \sum_{j=1}^{N} \max(0, 1 + \mathbf{w}_{r_{ij}}^T \mathbf{x}_{ij} - \mathbf{w}_{z_{ij}}^T \mathbf{x}_{ij}) \\ s.t. \quad & \text{if } L_i = 1, \sum_{\mathbf{x}_{ij} \in \mathbf{B}_i} z_{ij} > 0, \text{ and if } L_i = 0, z_{ij} = 0, \end{aligned} \tag{2.24}$$

where $r_{ij} = \arg\max_{c \in \{0, \cdots, C\}, c \notin z_{ij}} \mathbf{w}_c^T \mathbf{x}_{ij}$ and α is a scaling factor to promote the classi-

fication margin. Specifically, in (2.24), z_{ij} and r_{ij} are the indices of $\mathbf{x}_{ij}$ corresponding to the most confident (with largest decision value) and second most confident classification vectors in $\mathbf{W}$, respectively. The second term in (2.24) tries to maximize the classification margin between two most confident SVM classifiers and thus promotes the discriminativeness of the estimated classifier and induces the name "Max-Margin Dictionary Learning."

MMDL is optimized iteratively between steps of sampling a subset of training data based on each instance's "positiveness"; learning SVM classifiers, $\mathbf{W}$, using coordinate descent [3]; update of "positiveness" for each instance according to a sigmoid function; and a re-assignment of z_{ij} for each instance. Then the estimated SVM classifiers, $\mathbf{W}$, was adopted as the codebook and each image was represented as a distribution over the codebook using spatial pyramid matching [99]. Finally another linear SVM was trained for bag-level classification.

2.2.5 Other Multiple Instance Classifier Learning Algorithms

Beside the above reviewed MIL classifier learning algorithms, MILIS [100] alternates between the selection of an instance per bag as a prototype that represents its bag and training a linear SVM on these prototypes. MissSVM [101] solves the MIL problem using a semi-supervised SVM with the constraint that at least one point from each positive bag must be classified as positive. Recent approaches [69, 102–106] provide insightful and constructive view in MIL. Especially, Hoffman *et al.* [104] jointly exploit the image-level and bounding box labels and achieve state-of-the-art results in object detection. Li and Vasconcelos [105] further investigate MIL problem with labeling noise in negative bags and use "top instances" as the representatives of "soft bags", then proceed bag-level classification via latent-SVM [3].

Chapter 3

Previously Proposed Multiple Instance Concept Learning Algorithms

In the beginning of this chapter, two Functions of Multiple Instances (FUMI) approaches, extended FUMI (eFUMI) [67–70] and Dictionary Learning using FUMI (DL-FUMI) [71, 72], for learning representative target and non-target concepts are reviewed. Then, the discriminative target concept learning methods, multiple instance spectral matched filter (MI-SMF) and multiple instance adaptive cosine estimator (MI-ACE) [73] are investigated and discussed.

3.1 Extended Function of Multiple Instances

FUMI approaches [107, 108] assume each data point is some functional form of the concepts (or dictionary atoms) and tries to learn the unique features existing only in the positive bags. In particular, eFUMI extends FUMI to be able to learn target and non-target concepts

given only bag level labels for grouped training data indicating whether some proportion of target exist. It addresses this problem by assuming a set of latent variables that account for the true labels of each instances from positively labeled bags. eFUMI treats each data point as a convex combination of target and/or non-target concepts and learns a set of target concepts that are representative and unique features of the positive class and a set of non-target concepts that are a good generalization of the negative bags as well as false positive instances in the positive bags.

3.1.1 eFUMI

The goal of eFUMI is to estimate a target concept, $\mathbf{d}_T$, non-target concepts, $\mathbf{d}_k,\ \forall k = 1, \dots M$, the number of needed non-target concepts, M, and the abundances, $\mathbf{a}_j$, which define the convex combination of the concepts for each data point $\mathbf{x}_j$. The proposed objective function for learning these unknown variables is shown in (3.2). There are four terms in this objective function. The first term computes the squared error between the input data and its estimate found using the current target and non-target signatures and proportions where u is parameter constant controlling the relative importance of the first, second and third terms. The scaling value for $w_{l(\mathbf{x}_j)}$ is shown in (3.1),

$$w_{l(\mathbf{x}_j)} = \begin{cases} 1 & \text{if } l(\mathbf{x}_j) = 0 \\ \frac{\alpha N^+}{N^-} & \text{if } l(\mathbf{x}_j) = 1 \end{cases} \tag{3.1}$$

This scaling factor balances the influence between the positively and negatively labeled data. For example, if the parameter α is set to 1, then the weights on the target points are scaled such that the positive points has the same influence on the first term as the points

$$F = \frac{1}{2}(1-u)\sum_{j=1}^{N} w_j \left\| (\mathbf{x}_j - z_j a_{jT}\mathbf{d}_T - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k) \right\|_2^2 + \frac{u}{2}\sum_{k=T,1}^{M} \left\| \mathbf{d}_k - \boldsymbol{\mu}_0 \right\|_2^2 + \sum_{k=1}^{M}\gamma_k \sum_{j=1}^{N} a_{jk} \tag{3.2}$$

$$\begin{aligned} E[F] \;=\; & \sum_{z_j\in\{0,1\}} \left[\frac{1}{2}(1-u)\sum_{j=1}^{N} w_j P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)}) \left\| \mathbf{x}_j - z_j a_{jT}\mathbf{d}_T - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k \right\|_2^2 \right] \\ & + \frac{u}{2}\sum_{k=T,1}^{M} \|\mathbf{d}_k - \boldsymbol{\mu}_0\|_2^2 + \sum_{k=1}^{M}\gamma_k \sum_{j=1}^{N} a_{jk} \end{aligned} \tag{3.3}$$

from negative bags.

The second and third terms of the objective encourages target and non-target signatures that provide a tight fit around the data by minimizing the squared difference between each signature and the global data mean, $\boldsymbol{\mu}_0$. These terms were motivated by the volume-related term in the SPICE [35] algorithm. The fourth term is a sparsity promoting term used to determine M, the number of non-target signatures needed to describe the input data where $\gamma_k = \frac{\Gamma}{\sum_{j=1}^{N} a_{jk}^{(t-1)}}$, Γ is a parameter constant that controls the degree sparsity is promoted. Higher values of Γ generally result in a smaller estimate M value. The $a_{jk}^{(t-1)}$ values are the proportion values estimated in the previous iteration of the algorithm. Thus, as a the proportions for a particular endmember decrease, the weight of its associated sparsity promoting term increases. This approach for estimating the number of background endmembers follows the approach presented by the SPICE algorithm [35].

In (3.2) there are a set of hidden, latent variables, $z_j, j = 1, \cdots, N$, accounting for the unknown instance-level labels $l(\mathbf{x}_j)$. To address the fact that the z_j values are unknown, the expected values of the log likelihood with respect to z_j is taken as shown in (3.3). In (3.3), $\boldsymbol{\theta}^t$ is the set of parameters estimated at iteration t and $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)})$ is the probability of

individual points containing any proportion of target or not. $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)})$ is determined given the parameter set estimated in the previous iteration and the constraints of the bag-level labels, L_i, as shown in (3.4),

$$P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)}) = \begin{cases} e^{-\beta\left\|\mathbf{x}_j - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k\right\|_2^2} & \text{if } z_j = 0, L_i = 1 \\ 1 - e^{-\beta\left\|\mathbf{x}_j - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k\right\|_2^2} & \text{if } z_j = 1, L_i = 1 \\ 0 & \text{if } z_j = 1, L_i = 0 \\ 1 & \text{if } z_j = 0, L_i = 0 \end{cases} \tag{3.4}$$

where β is a scaling parameter and $r_j = \left\|\mathbf{x}_j - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k\right\|_2^2$ is the approximation residual between $\mathbf{x}_j$ and its representation using only background endmembers. The definition of $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)})$ in (3.6) indicates that if a point $\mathbf{x}_j$ is a nontarget point it should be fully represented by the background endmembers with very small residual r_j, and thus $P(z_j = 0|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)}) = e^{-\beta r_j} \rightarrow 1$. Otherwise, if $\mathbf{x}_j$ is a target point, it may not be well represented by only the background endmembers, so the residual r_j must be large and $P(z_j = 1|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)}) = 1 - e^{-\beta r_j} \rightarrow 1$. Note, z_j is unknown only for the positive bags; in the negative bags, z_j is fixed to 0. This constitutes the *E-step* of the EM algorithm.

The *M-step* is performed by optimizing (3.3) for each of the desired parameters. The method is summarized in Alg. 1.

3.1.2 eFUMI Optimization

In this section, the derivation of eFUMI update equations is provided. In order to solve for the update equation for the proportion values, $\mathbf{A}$, it should be pointed out that solving for the proportion value of one point is not dependent on any other points. Considering only

Algorithm 1 eFUMI EM algorithm

Initialize $\boldsymbol{\theta}^0 = \{\mathbf{d}_T, \mathbf{D}, \mathbf{A}\}$, $t = 1$
repeat
    E-step: Compute $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)})$ given $\boldsymbol{\theta}^{t-1}$
    M-step:
    Update $\mathbf{d}_T$ and $\mathbf{D}$ by maximizing (3.3) wrt. $\mathbf{d}_T$, $\mathbf{D}$
    Update $\mathbf{A}$ by maximizing (3.3) wrt. $\mathbf{A}$ s.t. the sum-to-one and non-negative constraints in Eq. (1.2)
    Prune each $\mathbf{d}_k, k = 1, \ldots, M$ if $\max_j(a_{jk}) \leq \tau$ where τ is a fixed threshold (e.g. $\tau = 10^{-6}$)
    $t \leftarrow t + 1$
until Convergence
return $\mathbf{d}_T, \mathbf{D}, \mathbf{A}$

points in positive bags, the eFUMI objective function becomes the form shown in (3.5) and a Lagrange multiplier term for the sum-to-one constraint is added in.

$$\begin{aligned} F^+ = & \sum_{j=1}^{N^+} \left[P(z_j = 0) \frac{1}{2}(1-u) w_j \left\| (\mathbf{x}_j - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k) \right\|_2^2 + P(z_j = 1) \frac{1}{2}(1-u) w_j \left\| (\mathbf{x}_j - a_{jT}\mathbf{d}_T - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k \right\|_2^2 \right] \\ & + \frac{u}{2} \sum_{k=1}^{M} \|\mathbf{d}_k - \boldsymbol{\mu}_0\|_2^2 + \frac{u}{2} \|\mathbf{d}_T - \boldsymbol{\mu}_0\|_2^2 + \sum_j \lambda_j^+ (a_{jT} + \sum_{k=1}^{M} a_{jk} - 1) + \sum_{k=1}^{M} \gamma_k \sum_{j=1}^{N^+} a_{jk} \end{aligned} \tag{3.5}$$

Then take partial derivative of (3.5) with respect to a_{jT} and a_{jk}, respectively.

$$\frac{\partial F^+}{\partial a_{jT}} = P(z_j = 1)(1-u) w_j (-1) \mathbf{d}_T^T (\mathbf{x}_j - a_{jT}\mathbf{d}_T - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k) + \lambda_j^+$$

$$\begin{aligned} \frac{\partial F^+}{\partial a_{jk}} = & P(z_j = 0)(1-u) w_j (-1) \mathbf{d}_k^T (\mathbf{x}_j - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k) \\ & + P(z_j = 1)(1-u) w_j (-1) \mathbf{d}_k^T (\mathbf{x}_j - a_{jT}\mathbf{d}_T - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k) + \lambda_j^+ + \gamma_k \end{aligned}$$

Let us denote $\alpha = (1-u)w_j(-1)$. Then, rewrite the above two functions into consistent matrix form and set the expression to 0.

$$\begin{aligned}\frac{\partial F^+}{\partial \mathbf{a}_j^+} =& \alpha P(z_j=0)\begin{bmatrix}0 & \mathbf{D}^-\end{bmatrix}^T(\mathbf{x}_j - \begin{bmatrix}0 & \mathbf{D}^-\end{bmatrix}\mathbf{a}_j^+) + \alpha P(z_j=1)\mathbf{D}^T(\mathbf{x}_j - \mathbf{D}\mathbf{a}_j^+) + \lambda_j^+\mathbf{1}_{(M+1)\times 1} + \begin{bmatrix}0\\ \mathbf{V}\end{bmatrix} \\ =& \Big[\alpha P(z_j=0)\begin{bmatrix}0 & \mathbf{D}^-\end{bmatrix}^T + \alpha P(z_j=1)\mathbf{D}^T\Big]\mathbf{x}_j \\ & - \Big\{\alpha P(z_j=0)\begin{bmatrix}0 & \mathbf{D}^-\end{bmatrix}^T\begin{bmatrix}0 & \mathbf{D}^-\end{bmatrix} + \alpha P(z_j=1)\mathbf{D}^T\mathbf{D}\Big\}\mathbf{a}_j^+ + \lambda_j^+\mathbf{1}_{(M+1)\times 1} + \begin{bmatrix}0\\ \mathbf{V}\end{bmatrix} = 0\end{aligned} \tag{3.6}$$

where $\mathbf{a}_j^+ = \begin{bmatrix}a_{jT}\\ a_{j1}\\ a_{j2}\\ \vdots\\ a_{jM}\end{bmatrix} = \begin{bmatrix}a_{jT}\\ \mathbf{a}_j^-\end{bmatrix}$, $\mathbf{V} = \begin{bmatrix}\gamma_1\\ \gamma_2\\ \vdots\\ \gamma_k\end{bmatrix}$, and $\mathbf{D} = \begin{bmatrix}\mathbf{d}_T & \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_M\end{bmatrix} = \begin{bmatrix}\mathbf{d}_T & \mathbf{D}^-\end{bmatrix}$.

$\mathbf{D}$ is the endmember matrix whose column corresponds to an endmember spectrum. $\mathbf{D}^-$ is a subset of $\mathbf{D}$ which accounts for constituent background endmembers. Similarly $\mathbf{a}_j^+$ is proportion vector for point $\mathbf{x}_j^+$ and $\mathbf{a}_j^-$ is a subset of $\mathbf{a}_j^+$, which accounts for the proportion values with respect to background endmembers. For points $\mathbf{x}_j^-$ from negative bags, a_{jT} is constrained to 0, so $\mathbf{a}_j = \begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}$.

Then, solve for $\mathbf{a}_j^+$,

$$\begin{aligned}\mathbf{a}_j^+ =& \Big\{P(z_j=0)\begin{bmatrix}0 & \mathbf{D}^-\end{bmatrix}^T\begin{bmatrix}0 & \mathbf{D}^-\end{bmatrix} + P(z_j=1)\mathbf{D}^T\mathbf{D}\Big\}^{-1}\cdot \\ & \left\{\Big[P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T + P(z_j=1)\mathbf{D}^T\Big]\mathbf{x}_j + \mathbf{1}_{(M+1)\times 1}\frac{\lambda_j^+}{\alpha} + \frac{1}{\alpha}\begin{bmatrix}0\\ \mathbf{V}\end{bmatrix}\right\}\end{aligned} \tag{3.7}$$

In order to enforce the sum-to-one constraint, multiply $\mathbf{1}_{1\times(M+1)}$ on both side of (3.7) and use the sum to one constraint $\mathbf{1}_{1\times(M+1)}\mathbf{a}_j^+ = 1$ to solve λ_j^+ shown as (3.9).

$$\begin{aligned}\lambda_j^+ &= \alpha\left(1-\mathbf{1}_{1\times(M+1)}\left\{P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}+P(z_j=1)\mathbf{D}^T\mathbf{D}\right\}^{-1}\right. \\ &\quad \left.\cdot\left\{\left[P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T+P(z_j=1)\mathbf{D}^T\right]\mathbf{x}_j+\frac{1}{\alpha}\begin{bmatrix}0\\ \mathbf{V}\end{bmatrix}\right\}\right) \\ &\quad \left\{\mathbf{1}_{1\times(M+1)}\left\{P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}+P(z_j=1)\mathbf{D}^T\mathbf{D}\right\}^{-1}\mathbf{1}_{(M+1)\times 1}\right\}^{-1}\end{aligned} \tag{3.8}$$

Substitute Eq. (3.9) to (3.7), finally, the update equation for the proportion values is shown in (3.9). Here, it is difficult to write $\mathbf{a}_j^+$ in matrix form because in $\left\{P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}+P(z_j=1)\mathbf{D}^T\mathbf{D}\right\}^{-1}$, $\mathbf{a}_j^+$ is related to $P(z_j)$ in an inverse matrix. So the proportion of points from positive bags must be updated point by point.

$$\begin{aligned}\mathbf{a}_j^+ &= \left\{P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}+P(z_j=1)\mathbf{D}^T\mathbf{D}\right\}^{-1} \\ &\quad \cdot\left\{\left[P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T+P(z_j=1)\mathbf{D}^T\right]\mathbf{x}_j+\frac{1}{\alpha}\begin{bmatrix}0\\ \mathbf{V}\end{bmatrix}\right. \\ &\quad +\mathbf{1}_{(M+1)\times 1}\cdot\left(1-\mathbf{1}_{1\times(M+1)}\left\{P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}+P(z_j=1)\mathbf{D}^T\mathbf{D}\right\}^{-1}\right. \\ &\quad \left.\cdot\left\{\left[P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T+P(z_j=1)\mathbf{D}^T\right]\mathbf{x}_j+\frac{1}{\alpha}\begin{bmatrix}0\\ \mathbf{V}\end{bmatrix}\right\}\right)\left(\mathbf{1}_{1\times(M+1)}\left\{P(z_j=0)\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}^T\right.\right. \\ &\quad \left.\left.\begin{bmatrix}\mathbf{0}_{n\times 1} & \mathbf{D}^-\end{bmatrix}+P(z_j=1)\mathbf{D}^T\mathbf{D}\right\}^{-1}\mathbf{1}_{(M+1)\times 1}\right)^{-1}\Bigg\}\end{aligned} \tag{3.9}$$

A similar derivation can be followed for points from negative bags (by simply excluding the term for the target endmember). The resulting update equation for negative points is

shown in (3.10).

$$\begin{aligned}\mathbf{A}^- &= (\mathbf{D}^{-T}\mathbf{D}^-)^{-1}\left[\mathbf{D}^{-T}\mathbf{X}^- - \frac{1}{1-u}\mathbf{R}\,(\mathbf{V})_{1\times N^-}\right. \\ &\left. +\mathbf{1}_{M\times 1}\cdot\frac{1-\mathbf{1}_{1\times M}(\mathbf{D}^{-T}\mathbf{D}^-)^{-1}\left(\mathbf{D}^{-T}\mathbf{X}^- - \frac{1}{1-u}\mathbf{R}\,(\mathbf{V})_{1\times N^-}\right)}{\mathbf{1}_{1\times M}(\mathbf{D}^{-T}\mathbf{D}^-)^{-1}\mathbf{1}_{M\times 1}}\right]\end{aligned} \tag{3.10}$$

To solve for the update for the endmember matrix $\mathbf{D}$, split objective function into two parts according to points from positive bags and points from negative bags and drop terms that are irrelevant to $\mathbf{D}$.

$$\begin{aligned}F &= \sum_{j=1}^{N^+}\left[P(z_j=0)\frac{1}{2}(1-u)w_j\left\|(\mathbf{x}_j-\sum_{k=1}^{M}a_{jk}\mathbf{d}_k)\right\|_2^2 + P(z_j=1)\frac{1}{2}(1-u)w_j\left\|(\mathbf{x}_j-a_{jT}\mathbf{d}_T-\sum_{k=1}^{M}a_{jk}\mathbf{d}_k\right\|_2^2\right] \\ &+\frac{1}{2}(1-u)\sum_{j=1}^{N^-}\left[\left\|(\mathbf{x}_j-\sum_{k=1}^{M}a_{jk}\mathbf{d}_k)\right\|_2^2\right] + \frac{u}{2}\sum_{k=1}^{M}\left\|(\mathbf{d}_k-\boldsymbol{\mu}_0)\right\|_2^2 + \frac{u}{2}\left\|(\mathbf{d}_T-\boldsymbol{\mu}_0)\right\|_2^2\end{aligned} \tag{3.11}$$

Then take the partial derivative of (3.11) with respect to $\mathbf{d}_T$ and $\mathbf{d}_k$, respectively.

$$\frac{\partial F}{\partial \mathbf{d}_T} = \sum_{j=1}^{N^+}\left[P(z_j=1)(-1)(1-u)w_j a_{jT}(\mathbf{x}_j - a_{jT}\mathbf{d}_T - \sum_{k=1}^{M}a_{jk}\mathbf{d}_k)\right] + u(\mathbf{d}_T-\boldsymbol{\mu}_0) \tag{3.12}$$

$$\begin{aligned}\frac{\partial F}{\partial \mathbf{d}_k} &= \sum_{j=1}^{N^+}\left[P(z_j=0)(-1)(1-u)w_j a_{jk}(\mathbf{x}_j-\sum_{k=1}^{M}a_{jk}\mathbf{d}_k) + P(z_j=1)(-1)(1-u)w_j a_{jk}(\mathbf{x}_j - a_{jT}\mathbf{d}_T\right. \\ &\left. -\sum_{k=1}^{M}a_{jk}\mathbf{d}_k)\right] + (1-u)\sum_{j=1}^{N^-}\left[-a_{jk}(\mathbf{x}_j-\sum_{k=1}^{M}a_{jk}\mathbf{d}_k)\right] + u(\mathbf{d}_k-\boldsymbol{\mu}_0)\end{aligned} \tag{3.13}$$

These can then be combined into matrix form and set the expression to 0.

$$\frac{\partial F}{\partial \mathbf{D}} = \sum_{j=1}^{N^+}\left[P(z_j=0)(-1)(1-u)w_j(\mathbf{x}_j-\mathbf{D}\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix})\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}^T + P(z_j=1)(-1)(1-u)w_j(\mathbf{x}_j-\mathbf{D}\mathbf{a}_j)\mathbf{a}_j^T\right]$$
$$+(1-u)\sum_{j=1}^{N^-}\left[-(\mathbf{x}_j-\mathbf{D}\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix})\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}^T\right] + u(\mathbf{D}-\mathbf{R}\,(\boldsymbol{\mu}_0)_{1\times(M+1)}) = 0 \tag{3.14}$$

Finally, the update equation for $\mathbf{E}$ is shown in (3.15).

$$\mathbf{D} = \left\{(1-u)w_j\sum_{j=1}^{N^+}\left[P(z_j=0)\mathbf{x}_j\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}^T + P(z_j=1)\mathbf{x}_j\mathbf{a}_j^{+T}\right] + (1-u)\sum_{j=1}^{N^-}\left[\mathbf{x}_j\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}^T\right] + u\cdot\mathbf{R}\,(\boldsymbol{\mu}_0)_{1\times(M+1)}\right\}$$
$$\cdot\left\{(1-u)w_j\sum_{j=1}^{N^+}\left[P(z_j=0)\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}^T + P(z_j=1)\mathbf{a}_j^+\mathbf{a}_j^{+T}\right] + (1-u)\sum_{j=1}^{N^-}\left[\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}\begin{bmatrix}0\\ \mathbf{a}_j^-\end{bmatrix}^T\right] + u\right\}^{-1} \tag{3.15}$$

3.1.3 eFUMI Initialization and Parameter Settings

Initialization for eFUMI and parameters are determined using the following. Non-target signatures are initialized by using the vertex component analysis (VCA) algorithm [109] on all data in the negatively-labeled bags. Then, using these initial non-target signatures, the data in the positively-labeled bags are unmixed. The data point with the largest reconstruction error using the initial non-target signatures is set as the initial target signature, $\mathbf{d}_T$. All proportion values are initialized to $\frac{1}{M+1}$ for all data points in positive bags and to $\frac{1}{M}$ for all points in negative bags (since the proportion on the target endmembers is fixed to 0).

There are a few parameters that must be set in eFUMI. The u parameters trades-off between the residual error term and the volume-related terms in the objective function. In all of our results, $u \in [0.01, 0.1]$. Valid values of u are in the set $(0, 1)$. Values that tend to 1 are appropriate for data with large noise levels. In contrast, values of u that tend towards 0 are appropriate for data with low noise levels and/or simulated data.

The initial value of M does not play a large role in accuracy of the algorithm provided that it is initialized to a value larger than needed number of non-target signatures. However, initializing with a very large M value will require a larger number of iterations in which to prune the unnecessary non-target signatures resulting in a longer running time. The sparsity promoting parameter Γ controls the degree of sparsity and the resulting number of non-target signatures. A larger Γ values tends to result in a smaller number of non-target signatures.

The scaling parameter, β, used in the calculation of $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)})$ aids in separating target and non-target points in positively labeled bags. As can be seen in the definition of $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1)})$, setting β is related to the magnitude of the input data, the number of background endmembers, and the spectral similarity between the target and background endmembers. For example, larger data magnitude corresponds to a larger reconstruction error in general and, thus, a smaller β value is needed. Similarly, more background endmembers result, in general, with smaller reconstruction error and, thus, a larger β value is needed. In our experience, normalizing the input data such that each data point has unit norm greatly helps in the setting of this parameter. In our experiments with normalized data, $\beta \in [30, 50]$ has been found to work well.

3.2 Dictionary Learning using Function of Multiple Instances

The goal of DL-FUMI [71, 72] is to leverage the benefits of dictionary learning approaches for problems in which only imprecise multiple instance learning type labels are available. Compared with the eFUMI method, DL-FUMI advantages in the following: (1) adopts the

knowledge of discriminative dictionary learning to estimate multiple representative signatures to accounts for the variability in the positive class; (2) introduces the use of a linear mixing model instead of convex mixing model, which is more suitable for general MIL problems.

Compared with supervised dictionary learning algorithms in the literature, DL-FUMI is able to handle the labeling uncertainty existent in the training data, *i.e.*, addresses the MIL problems. Furthermore, quite different from the majority of discriminative dictionary learning methods that estimate separate dictionaries for each class, DL-FUMI introduces a shared background model that also learns non-target concept from the positive bags. The advantage of this model over the class-specific dictionary learning is that the estimated target atoms only represents the unique characteristics of target resulting in improved target characterization and discrimination.

3.2.1 DL-FUMI

DL-FUMI models each instance as a sparse linear combination of target and/or background atoms $\mathbf{D}$, $\mathbf{x}_j \approx \mathbf{D}\mathbf{a}_j$, where $\mathbf{a}_j$ is the sparse vector of weights for instance $\mathbf{x}_j$. Positive bags (*i.e.*, $\mathbf{B}_i$ with $L_i = 1$, denoted as $\mathbf{B}_i^+$) contain at least one instance composed of some target:

$$\begin{aligned} &\text{if } L_i = 1, \exists \mathbf{x}_j \in \mathbf{B}_i^+ \text{ s.t.} \\ &\mathbf{x}_j = \sum_{t=1}^{T} a_{jt}\mathbf{d}_t^+ + \sum_{k=1}^{M} a_{jk}\mathbf{d}_k^- + \boldsymbol{\varepsilon}_j, a_{jt} \neq 0, \end{aligned} \tag{3.16}$$

where $\boldsymbol{\varepsilon}_j$ is a noise term. However, the number of instances in a positive bag with a target component is unknown.

If $\mathbf{B}_i$ is a negative bag (*i.e.*, $L_i = 0$, denoted as $\mathbf{B}_i^-$), then this indicates that $\mathbf{B}_i^-$ does not contain any target:

$$\text{if } L_i = 0, \forall \mathbf{x}_j \in \mathbf{B}_i^-, \mathbf{x}_j = \sum_{k=1}^{M} a_{jk}\mathbf{d}_k^- + \boldsymbol{\varepsilon}_j \tag{3.17}$$

Given this data mixing formulation, the goal of DL-FUMI is to estimate the dictionary $\mathbf{D} = \begin{bmatrix}\mathbf{D}^+ & \mathbf{D}^-\end{bmatrix} \in \mathbb{R}^{n\times(T+M)}$, where $\mathbf{D}^+ = \left[\mathbf{d}_1^+, \cdots, \mathbf{d}_T^+\right]$ are the T target atoms and $\mathbf{D}^- = \left[\mathbf{d}_1^-, \cdots, \mathbf{d}_M^-\right]$ are the M background atoms. This is accomplished by minimizing the objective function shown in Eq. (3.18), where $\mathbf{a}_i^+$ and $\mathbf{a}_i^-$ are subsets of $\mathbf{a}_i$ corresponding to $\mathbf{D}^+$ and $\mathbf{D}^-$, respectively.

$$F = \frac{1}{2}\sum_{j=1}^{N} w_j \left\| (\mathbf{x}_j - z_j \sum_{t=1}^{T} a_{jt}^+ \mathbf{d}_t^+ - \sum_{k=1}^{M} a_{jk}^- \mathbf{d}_k^-) \right\|_2^2 + \lambda \sum_{j=1}^{N} w_j \left\| \begin{bmatrix} z_j \mathbf{a}_j^+ \\ \mathbf{a}_j^- \end{bmatrix} \right\|_1 + \sum_{k=1}^{M}\sum_{t=1}^{T} \gamma_{kt} \langle \mathbf{d}_k^-, \mathbf{d}_{t^{\text{old}}}^+ \rangle \tag{3.18}$$

$$\begin{aligned} E[F] \quad = \quad & \sum_{z_j \in \{0,1\}} P(z_j | \mathbf{x}_j, \boldsymbol{\theta}^{(l-1)}) \left[\frac{1}{2}\sum_{j=1}^{N} w_j \left\| \mathbf{x}_j - z_j \sum_{t=1}^{T} a_{jt}^+ \mathbf{d}_t^+ - \sum_{k=1}^{M} a_{jk}^- \mathbf{d}_k^- \right\|_2^2 + \lambda \sum_{j=1}^{N} w_j \left\| \begin{bmatrix} z_j \mathbf{a}_j^+ \\ \mathbf{a}_j^- \end{bmatrix} \right\|_1 \right] \\ & + \sum_{k=1}^{M}\sum_{t=1}^{T} \gamma_{kt} \langle \mathbf{d}_k^-, \mathbf{d}_{t^{\text{old}}}^+ \rangle \end{aligned} \tag{3.19}$$

The first term in (3.18) computes the squared residual error between each instance and its estimate using the dictionary. In this term, a set of hidden binary latent variables $\{z_j\}_{j=1}^N$ that indicate whether an instance is or is not a target (*i.e.*, $z_j = 1$ when $\mathbf{x}_j$ contains target) are introduced. For all points in negative bags, $z_j = 0$. For points in positive bags, the value of z_j is unknown. Also, a weight w_j is included where $w_j = 1$ if $\mathbf{x}_j \in \mathbf{B}_i^-$ and $w_j = \psi$ if $\mathbf{x}_j \in \mathbf{B}_i^+$ where ψ is a fixed parameter. This weight helps balance terms when there is a

large imbalance between the number of negative and positive instances.

The second term is an l_1 regularization term to promote sparse weights. It also includes the latent variables, z_j, to account for the uncertain presence of target in positive bags.

The third term is a robust penalty term that promotes discriminative target atoms (and inspired by a term presented in [110]). Instead of using a fixed penalty coefficient, we introduce an adaptive coefficient γ_{kt} defined as:

$$\gamma_{kt} = \Gamma \frac{\langle \mathbf{d}_k^-, \mathbf{d}_t^+ \rangle}{\|\mathbf{d}_k^-\| \|\mathbf{d}_t^+\|} = \Gamma \cos \theta_{kt}, \tag{3.5}$$

where θ_{kt} is the vector angle between the k^{th} background atom and the t^{th} target atom. Since $sign(\gamma_{kt}) = sign(\langle \mathbf{d}_k^-, \mathbf{d}_t^+ \rangle)$, this discriminative term is always positive and will add large penalty when $\mathbf{d}_k^-$ and $\mathbf{d}_t^+$ have similar shapes. Thus, this term encourages a discriminative dictionary by promoting background atoms that are orthogonal to target atoms. In implementation, γ_{kt} is updated once per iteration using $\mathbf{d}_{k^{old}}^-$ and $\mathbf{d}_{t^{old}}^+$ which are the dictionary values from the previous iteration.

3.2.2 DL-FUMI Optimization

Expectation-Maximization is used to optimize (3.18) and estimate $\mathbf{D}$. During optimization, the fact that many of the binary latent variables $\{z_j\}_{j=1}^N$ are unknown is addressed by taking the expected value of the objective function with respect to z_j as shown in (3.20). In (3.20), $\boldsymbol{\theta}^l = \left\{\mathbf{D}, \{\mathbf{a}_j\}_{j=1}^N\right\}$is the set of parameters estimated at iteration l and $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(l-1)})$ is the probability that each instance is or is not a true target instance. During the E-step of

each iteration, $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(l-1)})$ is computed as:

$$P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(l-1)}) = \begin{cases} e^{-\beta\left\|\mathbf{x}_j - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k^-\right\|_2^2} & \text{if } z_j = 0, L_i = 1 \\ 1 - e^{-\beta\left\|\mathbf{x}_j - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k^-\right\|_2^2} & \text{if } z_j = 1, L_i = 1 \\ 0 & \text{if } z_j = 1, L_i = 0 \\ 1 & \text{if } z_j = 0, L_i = 0 \end{cases} \tag{3.6}$$

where β is a fixed scaling parameter.

If $\mathbf{x}_j$ is a non-target instance, then it should be characterized by the background atoms well, thus $P(z_j = 0|\mathbf{x}_j, \boldsymbol{\theta}^{(l-1)}) \approx 1$. Otherwise, if $\mathbf{x}_j$ is a true target instance, it will not be characterized well using only the background atoms and $P(z_j = 1|\mathbf{x}_j, \boldsymbol{\theta}^{(l-1)}) \approx 1$.

Algorithm 2 DL-FUMI EM algorithm

Initialize $\boldsymbol{\theta}^0 = \left\{\mathbf{D}, \{\mathbf{a}_j\}_{j=1}^N\right\}$, $l = 1$
repeat
    E-step: Compute $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(l-1)})$
    M-step:
    Update $\mathbf{d}_t^+$ using (3.7), $\mathbf{d}_t^+ \leftarrow \frac{1}{\|\mathbf{d}_t^+\|_2}\mathbf{d}_t^+, t = 1, \cdots, T$
    Update $\mathbf{d}_k^-$ using (3.8), $\mathbf{d}_k^- \leftarrow \frac{1}{\|\mathbf{d}_k^-\|_2}\mathbf{d}_k^-, k = 1, \cdots, M$
    Update $\{\mathbf{a}_j\}_{j=1}^{N^+}$ for $\mathbf{x}_j \in \mathbf{B}_i^+$ using gradient descent according to (3.10), (3.11)
    Update $\{\mathbf{a}_j\}_{j=1}^{N^-}$ for $\mathbf{x}_j \in \mathbf{B}_i^-$ using gradient descent according to (3.13)
    $l \leftarrow l + 1$
until Convergence
return $\mathbf{D}$, $\{\mathbf{a}_j\}_{j=1}^N$

The *M-step* is performed by iteratively optimizing (3.20) for each of the desired parameters. The dictionary $\mathbf{D}$ is updated atom-by-atom using a block coordinate descent scheme [111, 112]. The sparse weights, $\{\mathbf{a}_i\}_{i=1}^N$, are updated using an iterative shrinkage-thresholding algorithm [113, 114]. The method is summarized in Alg. 2 and the derivation of update equations are described as follows.

Similar to eFUMI, when updating the dictionary $\mathbf{D}$, the sparse weights $\{\mathbf{a}_j\}_{j=1}^N$ are held fixed. To update one of the atoms in $\mathbf{D}$, (3.20) is minimized with respect to the corresponding atom while keeping all other atoms constant. The resulting update equations for $\mathbf{d}_t^+$ and $\mathbf{d}_k^-$ are shown in (3.7) and (3.8).

$$\mathbf{d}_t^+ = \frac{\sum_{j=1}^{N+}\left[P(z_j=1)a_{jt}(\mathbf{x}_j - \sum_{l=1,l\neq t}^{T} a_{jl}\mathbf{d}_l^+ - \sum_{k=1}^{M} a_{jk}\mathbf{d}_k^-)\right]}{\sum_{j=1}^{N+}\left[P(z_j=1)a_{jt}^2\right]} \tag{3.7}$$

$$\begin{aligned}\mathbf{d}_k^- \quad = \quad & \left\{\sum_{j=1}^{N^+}\left[P(z_j=1)\psi a_{jk}(\mathbf{x}_j - \sum_{t=1}^{T} a_{jt}\mathbf{d}_t^+ - \sum_{l=1,l\neq k}^{M} a_{jl}\mathbf{d}_l^-) + P(z_j=0)\psi a_{jk}(\mathbf{x}_j - \sum_{l=1,l\neq k}^{M} a_{jl}\mathbf{d}_l^-)\right]\right. \\ & \left. + \sum_{j=1}^{N^-}\left[a_{jk}(\mathbf{x}_j - \sum_{l=1,l\neq k}^{M} a_{jl}\mathbf{d}_l^-)\right] - \Gamma\sum_{t=1}^{T}\cos\theta_{kt}\mathbf{d}_{told}^+\right\}\left\{\sum_{j=1}^{N+}\psi a_{jk}^2 + \sum_{j=1}^{N-}a_{jk}^2\right\}^{-1}\end{aligned} \tag{3.8}$$

Note, $P(z_j|\mathbf{x}_j, \boldsymbol{\theta}^{(t-1))}$ is denoted as $P(z_j)$ for simplicity.

When updating the sparse weights, $\{\mathbf{a}_j\}_{j=1}^N$, it should be noted that the sparse weight vector $\mathbf{a}_j$ for instance $\mathbf{x}_j$ is not dependent on any other instances.

For points from positively labeled bag, the gradient with respect to $\mathbf{a}_j$ without considering the l_1 penalty term is:

$$\begin{aligned}\frac{\partial F}{\partial \mathbf{a}_j} = & -\left[P(z_j=1)\mathbf{D}^+ \quad \mathbf{D}^-\right]^T\mathbf{x}_j + \left(P(z_j=1)\mathbf{D}^T\mathbf{D}\right. \\ & \left. + P(z_j=0)\left[\mathbf{0}_{d\times T} \quad \mathbf{D}^-\right]^T\left[\mathbf{0}_{d\times T} \quad \mathbf{D}^-\right]\right)\mathbf{a}_j.\end{aligned} \tag{3.9}$$

Then $\mathbf{a}_j$ at l^{th} iteration can be updated using gradient descent,

$$\mathbf{a}_j^l \leftarrow \mathbf{a}_j^{l-1} - \delta_j\frac{\partial F}{\partial \mathbf{a}_j^{l-1}}, \tag{3.10}$$

followed by a soft-thresholding:

$$\begin{cases} \mathbf{a}_j^{+l} = S_{\lambda P(z_j=1)}\left(\mathbf{a}_j^{+l}\right) \\ \mathbf{a}_j^{-l} = S_{\lambda}\left(\mathbf{a}_j^{-l}\right) \end{cases}, \tag{3.11}$$

where $S_\lambda\left(\mathbf{a}[k]\right) = sign(\mathbf{a}[k])\max(|\mathbf{a}[k]| - \lambda, 0)$, $k = 1, ..., n$.

Following a similar proof to that in [115], when the step length satisfies (3.12), the update of $\mathbf{a}_j$ using a gradient descent method with step length δ_i monotonically decreases the value of the objective function, where $Eig_{max}(\mathbf{P})$ denotes the maximum eigenvalue of $\mathbf{P}$. For simplicity, δ_j was set to $\frac{1}{Eig_{max}(\mathbf{D}^T\mathbf{D})}$ for all $\mathbf{a}_j$, $\mathbf{x}_j \in \mathbf{B}_i^+$.

$$\delta_j \in \left(0, \left(Eig_{max}\left(P(z_j = 0)\left[\mathbf{0}_{d\times T}\ \mathbf{D}^-\right]^T\left[\mathbf{0}_{d\times T}\ \mathbf{D}^-\right] + P(z_j = 1)\mathbf{D}^T\mathbf{D}\right)\right)^{-1}\right) \tag{3.12}$$

A similar update can be used for points from negative bags. The resulting update equation for negative points is:

$$\mathbf{a}_j^l \leftarrow S_\lambda\left(\mathbf{a}_j^{l-1} + \frac{1}{Eig_{max}\left(\mathbf{D}^{-T}\mathbf{D}^-\right)}\left(\mathbf{D}^{-T}(\mathbf{x}_j - \mathbf{D}^-\mathbf{a}_j^{l-1})\right)\right) \tag{3.13}$$

The sparse weights corresponding to target dictionary atoms are set to 0 for all points in negative bags.

3.2.3 Classification using Estimated Dictionary

Given $\mathbf{D}$, a confidence that the j^{th} instance is target can be computed using a ratio of the reconstruction errors given the target and background atoms, $\mathbf{D}$, vs. background atoms,

$\mathbf{D}^-$:

$$\Lambda_j = \frac{\|\mathbf{x}_j - \mathbf{p}_j\mathbf{D}^-\|^2}{\|\mathbf{x}_j - \mathbf{a}_j\mathbf{D}\|^2}, \tag{3.14}$$

where $\mathbf{p}_j$ are the sparse weights of the j^{th} instance given only the non-target atoms $\mathbf{D}^-$. If the numerator has a large error and the denominator has a low error, then the target atoms are needed to reconstruct instance $\mathbf{x}_j$.

3.3 Multiple Instance Spectral Matched Filter and Multiple Instance Adaptive Coherence/Cosine Detector

Quite different from FUMI algorithms that aim to recover exactly the target concepts mixed in the training data, the discriminative target concept learning tries to find a set of target concepts that help discriminate the target instances. Two algorithms, MI-SMF and MI-ACE [73], are reviewed in this section. In general, the discussed algorithms aim to maximize the detection statistics of some signature based detectors under the MIL problem definition. Experimental results show that the estimated target concepts may look quite different from the ground truth, but achieve better detection performance.

3.3.1 MI-SMF and MI-ACE

The MI-SMF and MI-ACE maximize the detector response of SMF and ACE under a MIL problem definition. Compared with FUMI algorithms that minimize the Euclidean distance between the input data and its reconstruction error using the estimated target and non-target concepts, MI-SMF and MI-ACE maximize the cosine similarity between the target concept

and the input instance which is found to be more robust to highly mixed, noisy training data and efficient in optimization.

Following the MIL definition, a positive bag (*i.e.*, $\mathbf{B}_i$ with $L_i = 1$, denoted as $\mathbf{B}_i^+$) must contain at least one instance contains some amount of target following the target present hypothesis shown in Eq. (3.15):

$$\text{if } L_i = 1, \exists \mathbf{x}_{ij} \in \mathbf{B}_i^+ \text{ s.t. } \mathbf{x}_{ij} \sim \mathcal{N}\left(a_{ij}\mathbf{s} + \boldsymbol{\mu}_b, \sigma_1^2 \Sigma_b\right), a_{ij} \neq 0. \tag{3.15}$$

If $\mathbf{B}_i$ is a negative bag (*i.e.*, $L_i = 0$, denoted as $\mathbf{B}_i^-$), then this indicates that $\mathbf{B}_i^-$ is composed of all non-target instances, following the target absent hypothesis:

$$\text{if } L_i = 0, \mathbf{x}_{ij} \sim \mathcal{N}\left(\boldsymbol{\mu}_b, \sigma_0^2 \Sigma_b\right) \forall \mathbf{x}_{ij} \in \mathbf{B}_i^- \tag{3.16}$$

Given this problem formulation, the proposed objective function for MI-SMF and MI-ACE is to learn the target signature, $\mathbf{s}$, is shown in Eg. (3.17). The objective function maximizes the detector response of at least one instances in each positive bag and minimizes the detector response over all negative instances.

$$\arg\max_{\mathbf{s}} \frac{1}{K^+} \sum_{i:L_i=1} \Lambda(\mathbf{x}_i^*, \mathbf{s}) - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \Lambda(\mathbf{x}_{ij}, \mathbf{s}), \tag{3.17}$$

where $\mathbf{x}_i^*$ is the selected representative instance from the positive bag B_i^+ with the largest detector response, $\Lambda(\cdot, \mathbf{s})$, given a target signature, $\mathbf{s}$:

$$\mathbf{x}_i^* = \arg \max_{\mathbf{x}_{ij} \in B_i^+} \Lambda(\mathbf{x}_{ij}, \mathbf{s}) \tag{3.18}$$

The assumption that each positive bag could be represented by the most positive instance in it inherits the advantages of the EM-DD paper [78].

In order to maximize Eq. (3.17) with respect to $\mathbf{s}$, first apply some transformation on the ACE detector:

$$\begin{aligned}
\Lambda_{ACE}(\mathbf{x},\mathbf{s}) &= \frac{\mathbf{s}^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{x}-\boldsymbol{\mu}_b)}{\sqrt{\mathbf{s}^T\boldsymbol{\Sigma}_b^{-1}\mathbf{s}}\sqrt{(\mathbf{x}-\boldsymbol{\mu}_b)^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{x}-\boldsymbol{\mu}_b)}} \\
&= \frac{\mathbf{s}^T\mathbf{U}\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{U}^T(\mathbf{x}-\boldsymbol{\mu}_b)}{\sqrt{\mathbf{s}^T\mathbf{U}\mathbf{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{s}}\sqrt{(\mathbf{x}-\boldsymbol{\mu}_b)^T\mathbf{U}\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{U}^T(\mathbf{x}-\boldsymbol{\mu}_b)}} \\
&= \left(\frac{\hat{\mathbf{s}}}{\|\hat{\mathbf{s}}\|}\right)^T\left(\frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|}\right) \\
&= \hat{\hat{\mathbf{s}}}^T\hat{\hat{\mathbf{x}}} \qquad (3.19)
\end{aligned}$$

where $\hat{\mathbf{x}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{U}^T(\mathbf{x}-\boldsymbol{\mu}_b)$, $\hat{\mathbf{s}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{s}$, $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors and eigenvalues of the background covariance matrix, $\boldsymbol{\Sigma_b}$, respectively, $\hat{\hat{\mathbf{s}}} = \frac{\hat{\mathbf{s}}}{\|\hat{\mathbf{s}}\|}$ and $\hat{\hat{\mathbf{x}}} = \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|}$. It can be clearly seen from Eq. (3.19) that the ACE detector response is the cosine value between a test data point, $\mathbf{x}$, and a target signature, $\mathbf{s}$ after whitening. Thus, the objective function (3.17) for MI-ACE can be written as:

$$\arg\max_{\hat{\mathbf{s}}} \frac{1}{K^+}\sum_{i:L_i=1}\hat{\hat{\mathbf{s}}}^T\hat{\hat{\mathbf{x}}}_i^* - \frac{1}{K^-}\sum_{i:L_i=0}\frac{1}{N_i^-}\sum_{\mathbf{x}_{ij}\in B_i^-}\hat{\hat{\mathbf{s}}}^T\hat{\hat{\mathbf{x}}}_{ij}, \text{ such that } \hat{\hat{\mathbf{s}}}^T\hat{\hat{\mathbf{s}}} = 1. \qquad (3.20)$$

The l_2 norm constraint, $\hat{\hat{\mathbf{s}}}^T\hat{\hat{\mathbf{s}}} = 1$, is resulted from the normalization term in Eq. (3.19). The optimum for (3.20) can be derived by solving the Lagrangian:

$$\hat{\hat{\mathbf{s}}} = \frac{\mathbf{t}}{\|\mathbf{t}\|}, \text{ where } \mathbf{t} = \frac{1}{K^+}\sum_{i:L_i=1}\hat{\hat{\mathbf{x}}}_i^* - \frac{1}{K^-}\sum_{i:L_i=0}\frac{1}{N_i^-}\sum_{\mathbf{x}_{ij}\in B_i^-}\hat{\hat{\mathbf{x}}}_{ij} \qquad (3.21)$$

Similarly, the objective function for MI-SMF can be written as:

$$\arg\max_{\hat{\hat{\mathbf{s}}}} \frac{1}{K^+} \sum_{i:L_i=1} \hat{\hat{\mathbf{s}}}^T \hat{\mathbf{x}}_i^* - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \hat{\hat{\mathbf{s}}}^T \hat{\mathbf{x}}_{ij}, \text{ such that } \hat{\hat{\mathbf{s}}}^T \hat{\hat{\mathbf{s}}} = 1. \tag{3.22}$$

resulting in the following update equation for $\hat{\hat{\mathbf{s}}}$:

$$\hat{\hat{\mathbf{s}}} = \frac{\mathbf{t}}{\|\mathbf{t}\|}, \text{ where } \mathbf{t} = \frac{1}{K^+} \sum_{i:L_i=1} \hat{\mathbf{x}}_i^* - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \hat{\mathbf{x}}_{ij}, \tag{3.23}$$

where it shows the difference between MI-SMF and MI-ACE that MI-SMF takes the unnormalized data for consideration, which is consistent with the SMF formulation shown in Eq. (1.5).

MI-SMF and MI-ACE alternates between two steps: (1) selecting representative instances from each positive bag and (2) updating the target concept $\mathbf{s}$. The algorithm is summarized in Alg. 3.

Algorithm 3 MI-SMF/MI-ACE

1: Compute $\boldsymbol{\mu_b}$ and $\boldsymbol{\Sigma}_b$ as the mean and covariance of all instances in the negative bags
2: Subtract the background mean and whiten all instances, $\hat{\mathbf{x}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu}_b)$
3: If MI-ACE, normalize: $\hat{\hat{\mathbf{x}}} = \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|}$
4: Initialize $\hat{\hat{\mathbf{s}}}$ using the instance in a positive bag resulting in largest objective function value
5: **repeat**
6: Update the selected instances, $\mathbf{x}_i^*$, for each positive bag, $\mathbf{B}_i^+$ using (3.18)
7: Update $\hat{\hat{\mathbf{s}}}$ using (3.21) for MI-ACE or (3.23) for MI-SMF
8: **until** Stopping Criterion Reached
9: **return** $\mathbf{s} = \frac{\mathbf{t}}{\|\mathbf{t}\|}$, where $\mathbf{t} = \mathbf{U}\mathbf{V}^{\frac{1}{2}}\hat{\hat{\mathbf{s}}}$

MI-SMF and MI-ACE stop when there is no change in the selection of instances from positive bags across two iterations. Similar to [78], since there exists a finite set of possible selection of positive instances given a finite training bags, the convergence of MI-SMF and

MI-ACE is guaranteed. In the experiments shown in this work, MI-SMF and MI-ACE generally converged with less than 7 iterations.

3.3.2 MI-ACE and MI-SMF Optimization

To derive the update equation for MI-ACE (updates for MI-SMF can be similarly derived), the Lagrangian for MI-ACE objective function in (3.20) is shown below:

$$\mathscr{L} = \frac{1}{K^+} \sum_{i:L_i=1} \hat{\hat{\mathbf{s}}}^T \hat{\hat{\mathbf{x}}}_i^* - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \hat{\hat{\mathbf{s}}}^T \hat{\hat{\mathbf{x}}}_{ij} - \lambda \left(\hat{\hat{\mathbf{s}}}^T \hat{\hat{\mathbf{s}}} - 1 \right) \tag{3.24}$$

where λ is the Lagrange multiplier. The derivative of the Lagrangian with respect to $\hat{\hat{\mathbf{s}}}$ is:

$$\frac{\partial \mathscr{L}}{\partial \hat{\hat{\mathbf{s}}}} = \frac{1}{K^+} \sum_{i:L_i=1} \hat{\hat{\mathbf{x}}}_i^* - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \hat{\hat{\mathbf{x}}}_{ij} - 2\lambda \hat{\hat{\mathbf{s}}} \tag{3.25}$$

We can then set (3.25) to zero and solve for $\hat{\hat{\mathbf{s}}}$:

$$\hat{\hat{\mathbf{s}}} = \frac{1}{2\lambda} \left(\frac{1}{K^+} \sum_{i:L_i=1} \hat{\hat{\mathbf{x}}}_i^* - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \hat{\hat{\mathbf{x}}}_{ij} \right) \tag{3.26}$$

Then, define $\mathbf{t}$ as:

$$\mathbf{t} = \frac{1}{K^+} \sum_{i:L_i=1} \hat{\hat{\mathbf{x}}}_i^* - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \hat{\hat{\mathbf{x}}}_{ij}. \tag{3.27}$$

To determine the value of the Lagrange multiplier, λ, we must determine the value for λ that enforces the constraint that $\hat{\hat{\mathbf{s}}}^T \hat{\hat{\mathbf{s}}} = 1$. Thus, $\lambda = \frac{\|\mathbf{t}\|}{2}$ which results in the final update

equation for $\hat{\hat{\mathbf{s}}}$:

$$\hat{\hat{\mathbf{s}}} = \frac{\mathbf{t}}{\|\mathbf{t}\|}, \text{ where } \mathbf{t} = \frac{1}{K^+} \sum_{i:L_i=1} \hat{\hat{\mathbf{x}}}_i^* - \frac{1}{K^-} \sum_{i:L_i=0} \frac{1}{N_i^-} \sum_{\mathbf{x}_{ij} \in B_i^-} \hat{\hat{\mathbf{x}}}_{ij}. \tag{3.28}$$

The derivation for the update equation for the MI-SMF target signature is identical to what is shown above except $\hat{\mathbf{x}}$ is used in place of $\hat{\hat{\mathbf{x}}}$ in all of the preceding equations in this section.

Chapter 4

Multiple Instance Hybrid Estimator

In this chapter, we present the proposed multiple instance hybrid estimator (MI-HE) [74, 75] framework. MI-HE is a discriminative target concept learning algorithm for problems with mixed data and label uncertainty. Quite different from the FUMI algorithms [69, 71] that learn representative concept from reconstruction error, MI-HE learns discriminative target concept that maximizes the detection performance. Specifically, the estimated target concept by MI-HE maximizes the detection response of structured hybrid detector (HSD) [56, 65] under a generalized mean model. Furthermore, the FUMI algorithms do not exploit the entire label information from the training data, *i.e.*, the FUMI algorithms combine all positive bags together into a big positive bag and thus discard the information that each positive bag must contain at least one positive instance. On the contrary, MI-HE adopts a generalized mean model to differentiate each individual positive bag. Experimental results show that the estimated target concept by MI-HE achieves better detection performance.

Compared with the existing MIL algorithms in the literature, MI-HE explicitly deals with the following difficulties in MIL and target detection problems:

1. **The number of negative training instances are often much more than that of positive training instances.** MI-HE addresses this problem by applying the hybrid detector only to the instances from the positively labeled bags. The negative bags are only needed to refine the background concepts.

2. **Mixed training data.** MI-HE explicitly utilizes a linear mixture model and maximizes the detection statistics. So the resultant target and background concepts applied to testing data can also perform sub-pixel detection.

3. **Learning discriminative target concepts.** Maximizing discrimination ability in MI-HE is accomplished in two ways: (1) the proposed MI-HE algorithm assumes a shared background concepts set between the positive and negative bags; and (2) an efficient discriminative term is adopted.

4.1 Multiple Instance Hybrid Estimator Learning Framework

MI-HE starts from the bag-level likelihood measurement and wants to maximizes the probability of the labels of the given bags,

$$J_1 = \prod_{i=1}^{K^+} \Pr(L_i = +|\mathbf{B}_i^+) \cdot \prod_{i=K^++1}^{K^++K^-} \Pr(L_i = -|\mathbf{B}_i^-). \tag{4.1}$$

Since the MIL problem states that there must be at least one positive instance in each positive bag and each negative bag must consist of only negative instances, we can approximate the probability of an individual bag to the instances in each bag, as shown in Eq. (4.2). Specifically, the probability for a positive bag to be positive is substituted by the instance

in this bag with highest "positiveness" and the probability for a negative bag to be negative is represented by the joint probability of all instances in this bag to be negative.

$$J_2 = \prod_{i=1}^{K^+} \max_{\mathbf{x}_{ij} \in \mathbf{B}_i^+} \Pr(l_{ij} = +|\mathbf{B}_i^+) \cdot \prod_{i=K^++1}^{K^++K^-} \prod_{j=1}^{N_i} \Pr(l_{ij} = -|\mathbf{B}_i^-). \tag{4.2}$$

Eq. (4.2) contains a $\max$ operation that is difficult to optimize numerically. Some algorithms in the literature [77, 78] adopt a noisy-OR model instead of using $\max$. However, experimental results show that the noisy-OR model is highly non-smooth and needs to be repeated with many different initializations (typically using every positive training instance) to avoid local optima. In the proposed approach, we adopt the generalized mean as an alternative of $\max$ operation, as shown in (4.3).

$$J_3 = \prod_{i=1}^{K^+} \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \Pr(l_{ij} = +|\mathbf{B}_i^+)^b \right)^{\frac{1}{b}} \cdot \prod_{i=K^++1}^{K^++K^-} \prod_{j=1}^{N_i} \Pr(l_{ij} = -|\mathbf{B}_i^-), \tag{4.3}$$

where $b \in [-\infty, +\infty]$ is a real number controlling the function to approximately vary from $\min$ to $\max$.

Then taking the negative logarithm and scaling the second term of Eq. (4.3) results in:

$$-\ln J_3 = -\sum_{i=1}^{K^+} \frac{1}{b} \ln \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \Pr(l_{ij} + = +|\mathbf{B}_i^+)^b \right) - \rho \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} \ln \Pr(l_{ij} = -|\mathbf{B}_i^-), \tag{4.4}$$

where the scaling factor ρ is usually set to be smaller than one to control the influence of negative bags.

Similar to DL-FUMI [71], here each instance is modeled as a sparse linear combination of target and/or background concepts $\mathbf{D}$, $\mathbf{x}_i \approx \mathbf{D}\mathbf{a}_i$, where $\mathbf{a}_i$ is the sparse vector of weights

for instance $\mathbf{x}_i$. Each positive bag contains at least one instance composed of some target:

$$\begin{aligned} &\text{if } L_i = 1, \exists \mathbf{x}_{ij} \in \mathbf{B}_i^+ \text{ s.t.} \\ &\mathbf{x}_{ij} = \sum_{t=1}^{T} a_{it}\mathbf{d}_t^+ + \sum_{k=T+1}^{T+M} a_{ik}\mathbf{d}_k^- + \boldsymbol{\varepsilon}_{ij}, a_{it} \neq 0, \end{aligned} \tag{4.5}$$

where $\boldsymbol{\varepsilon}_i$ is a noise term. Each negatively labeled bag $\mathbf{B}_i^-$ should not contain any target:

$$\text{if } L_i = 0, \forall \mathbf{x}_{ij} \in \mathbf{B}_i^-, \mathbf{x}_{ij} = \sum_{k=T+1}^{T+M} a_{ik}\mathbf{d}_k^- + \boldsymbol{\varepsilon}_{ij}. \tag{4.6}$$

Given the above data model, we introduce the hybrid detector to estimate if instances from positive bags are positive target points. Specifically, define the following term,

$$\Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+) = \exp\left(-\beta \frac{\|\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}\|^2}{\|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^2}\right), \tag{4.7}$$

where $\mathbf{D} = \begin{bmatrix}\mathbf{D}^+ & \mathbf{D}^-\end{bmatrix} \in \mathbb{R}^{n\times(T+M)}$, $\mathbf{D}^+ = \left[\mathbf{d}_1^+, \cdots, \mathbf{d}_T^+\right]$ is the set of T target concepts and $\mathbf{D}^- = \left[\mathbf{d}_{T+1}^-, \cdots, \mathbf{d}_{T+M}^-\right]$ is the set of M background concepts, β is a scaling parameter; $\mathbf{a}_{ij}$ and $\mathbf{p}_{ij}$ are the sparse representation of $\mathbf{x}_{ij}$ given entire concept set $\mathbf{D}$ and background concept set $\mathbf{D}^-$, $\mathbf{r}_{ij} = (\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij})$ and $\mathbf{q}_{ij} = (\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij})$ are the corresponding reconstruction residual vectors, respectively. Further, $\mathbf{a}_{ij} = [\mathbf{a}_{ij}^+;\ \mathbf{a}_{ij}^+]$, where $\mathbf{a}_{ij}^+$ and $\mathbf{a}_{ij}^-$ are subsets of $\mathbf{a}_{ij}$ corresponding to $\mathbf{D}^+$ and $\mathbf{D}^-$, respectively. Since $\mathbf{D}$ is a super set of $\mathbf{D}^-$, theoretically the reconstruction error of $\mathbf{x}_{ij}$ using $\mathbf{D}$ (the numerator) should be always smaller than that using $\mathbf{D}^-$ (the denominator). Specifically, solving the sparse representation $\mathbf{a}$ given a dictionary set $\mathbf{D}$ is modeled as the Lasso problem [116, 117] shown in Eq. (4.8):

$$\mathbf{a}^* = \arg\min \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1, \tag{4.8}$$

where λ is a scaling factor to control the sparsity of $\mathbf{a}$. The solving of l_1 regularized least squares have been investigated extensively in the literature [118–120]. Here we adopt the iterative shrinkage-thresholding algorithm (ISTA) [113, 114] for solving the sparse codes $\mathbf{a}$.

The definition of $\Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)$ in (4.7) indicates that if a point $\mathbf{x}_{ij} \in \mathbf{B}_i^+$ is a true positive point, it may not be well represented by only the non-target concepts, so the residual error approximated by the entire concepts, $\|\mathbf{r}_{ij}\|^2$, will be much smaller than that by the background concepts $\|\mathbf{q}_{ij}\|^2$, thus $\Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+) = \exp\left(-\beta \frac{\|\mathbf{r}_{ij}\|^2}{\|\mathbf{q}_{ij}\|^2}\right) \rightarrow 1$. Otherwise, if $\mathbf{x}_{ij} \in \mathbf{B}_i^+$ is a false positive point, $\|\mathbf{r}_{ij}\|^2 \approx \|\mathbf{q}_{ij}\|^2$, thus $\Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+) = \exp\left(-\beta \frac{\|\mathbf{r}_{ij}\|^2}{\|\mathbf{q}_{ij}\|^2} \rightarrow 0\right)$.

For points from negative bags, following Eq. (4.6), we model the reconstruction error of points $\mathbf{x}_{ij} \in \mathbf{B}_i^-$ as a zero mean Gaussian distribution, shown as Eq. (4.9),

$$\Pr(l_{ij} = -|\mathbf{B}_i^-) = \exp\left(-\|\mathbf{x}_{ij} - \mathbf{D}^- \mathbf{p}_{ij}\|^2\right), \tag{4.9}$$

where $\mathbf{p}_{ij}$ is the sparse representation of $\mathbf{x}_{ij}$ given $\mathbf{D}^-$ and is solved by Eq. (4.8). Here instead of applying the hybrid detector, we use a least squares to represent the residual error of $\mathbf{x}_{ij}$. This indicates that the negative points should be fully represented by just the non-target concepts, $\mathbf{D}^-$. The intuitive understanding of this assumption is that minimizing the least squares of all of the negative points provides a good description of the background. Moreover, because there are typically many more negative points in negatively labeled bags than true positive points in positive bags, the target concept estimation may be biased if the hybrid detector was also applied to negative instances.

Thus far, we have constructed an objective function that learns a set of concepts that maximize the hybrid sub-pixel detector statistics of the positive bags and characterize the negative bags. However, given the objective function so far, there is no guarantee that the

estimated target concept captures only the discriminative features of the positive class and is discriminative from the negative class. Inspired by the discriminative terms proposed by the Dictionary Learning with Structured Incoherence [110] and the Fisher Discrimination Dictionary Learning (FDDL) algorithm [121, 122], we propose a cross incoherence term $Q(\mathcal{X}, \mathbf{D}^+, \mathcal{A})$ shown in Eq. 4.10 to complete the objective, where $\mathcal{X}$ is the union of all instances from negatively labeled bags, $\mathbf{D}^+$ is the target concept set which is the subset of $\mathbf{D} = [\mathbf{D}^+ \ \mathbf{D}^-]$, $\mathcal{A} = [\mathcal{A}^+ \ \mathcal{A}^-]$ is the sparse codes matrix of $\mathcal{X}$ with respect to the entire concepts $\mathbf{D}$.

$$\begin{aligned} Q(\mathcal{X}, \mathbf{D}^+, \mathcal{A}) &= \frac{\alpha}{2} \| Diag\left((\mathbf{D}^+ \mathcal{A}^+)^T \mathcal{X} \right) \|_2^2 \\ &= \frac{\alpha}{2} \sum_{i=K^+ +1}^{K} \sum_{j=1}^{N_i} \left((\mathbf{D}^+ \mathbf{a}_{ij}^+)^T \mathbf{x}_{ij} \right)^2 \end{aligned} \tag{4.10}$$

The understanding of the proposed cross incoherence term is presented by examining the reconstruction of the negative data set $\mathcal{X}$. First of all, $\mathcal{X}$ should be well represented by the non-target concept set $\mathbf{D}^-$, *i.e.*, $\mathcal{X} \approx \mathbf{D}^- \mathbf{P}$. This is fulfilled by inclusion of the term in Eq. (4.9). Second, since $\mathbf{D} = [\mathbf{D}^+ \ \mathbf{D}^-]$ is a superset of $\mathbf{D}^-$, the reconstruction error of $\mathcal{X}$ by the entire concept set $\mathbf{D}$ is also small, *i.e.*, $\mathcal{X} \approx \mathbf{D}^+ \mathcal{A}^+ + \mathbf{D}^- \mathcal{A}^- = \mathbf{R}^+ + \mathbf{R}^-$. In order to have a target concept $\mathbf{D}^+$ that is distinct from the negative data, it is expected that the reconstruction of $\mathcal{X}$ with respect to the target concept, $\mathbf{R}^+$, should either maintain small energy or else have a bad representation of $\mathcal{X}$, and thus Eq. 4.10 is optimized.

The final objective function is shown in Eq. 4.11, which contains three terms: generalized mean (GM) term (first), background data fidelity term (second) and the cross incoherence (discriminative) term (third):

$$J_4 = -\sum_{i=1}^{K^+}\frac{1}{b}\ln\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\exp\left(-\beta\frac{\|\mathbf{x}_{ij}-\mathbf{D}\mathbf{a}_{ij}\|^2}{\|\mathbf{x}_{ij}-\mathbf{D}^-\mathbf{p}_{ij}\|^2}\right)^b\right)+\rho\sum_{i=K^++1}^{K^++K^-}\sum_{j=1}^{N_i}\|\mathbf{x}_{ij}-\mathbf{D}^-\mathbf{p}_{ij}\|^2 + \frac{\alpha}{2}\sum_{i=K^++1}^{K}\sum_{j=1}^{N_i}\left((\mathbf{D}^+\mathbf{a}_{ij}^+)^T\mathbf{x}_{ij}\right)^2, \quad (4.11)$$

4.2 Optimization

The optimization of Eq. (4.11) can be decomposed into two sub-problems, updating the concepts $\mathbf{D}$ and the sparse representation of $\mathbf{a}$ alternatively.

4.2.1 Concept Optimization

Similar to the Dictionary Learning using Singular Value Decomposition (K-SVD) approach [85], the optimization of target and background concepts is performed by taking gradient descent with respect to one atom at a time and holding the rest fixed. Denote f_{GM} as the generalized mean part of Eq. (4.11) and expend the logarithm of f_{GM}:

$$\begin{aligned} f_{GM} &= -\sum_{i=1}^{K^+}\frac{1}{b}\ln\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\Lambda(\mathbf{x}_{ij},\mathbf{D}|\mathbf{B}_i^+)^b\right) \\ &= -\sum_{i=1}^{K^+}\frac{1}{b}\ln\frac{1}{N_i}-\sum_{i=1}^{K^+}\frac{1}{b}\ln\left(\sum_{j=1}^{N_i}\Lambda(\mathbf{x}_{ij},\mathbf{D}|\mathbf{B}_i^+)^b\right) \end{aligned} \quad (4.12)$$

Remove the constant part in f_{GM} and take the partial derivative with respect to $\mathbf{d}$, here

$\mathbf{d}$ is a symbolic notation for any atom in $\mathbf{D}$, shown in (4.13):

$$\frac{\partial f_{GM}}{\partial \mathbf{d}} = -\sum_{i=1}^{K^+} \frac{1}{\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^b} \cdot \left(\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^{b-1} \cdot \frac{\partial \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)}{\partial \mathbf{d}} \right) \tag{4.13}$$

Then take the partial derivative on the fidelity (second) term of the objective function Eq. (4.11) with respect to the background concept, $\mathbf{d}_k^-$:

$$\begin{aligned} \frac{\partial - \rho \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} \ln \Pr(l_{ij} = -|\mathbf{B}_i^-)}{\partial \mathbf{d}_k^-} &= \frac{\partial \rho \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} \|\mathbf{x}_{ij} - \mathbf{D}^- \mathbf{p}_{ij}\|^2}{\partial \mathbf{d}_k^-} \\ &= \rho \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} -2p_{ijk}(\mathbf{x}_{ij} - \mathbf{D}^- \mathbf{p}_{ij}) \\ &= \rho \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} -2p_{ijk}\mathbf{q}_{ij}, \end{aligned} \tag{4.14}$$

where p_{ijk} is the kth element in $\mathbf{p}_{ij}$ corresponding to $\mathbf{d}_k^-$.

The partial derivative of the cross incoherence (third) term corresponding to $\mathbf{d}_t^+$ is:

$$\frac{\partial Q(\mathcal{X}, \mathbf{D}^+, \mathcal{A})}{\partial \mathbf{d}_t^+} = \alpha \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} (\mathbf{D}^+ \mathbf{a}_{ij}^+)^T \mathbf{x}_{ij} \cdot a_{ijt}^+ \mathbf{x}_{ij} \tag{4.15}$$

The partial derivatives of the negative objective function Eq. (4.11) with respect to $\mathbf{d}_t^+$ and $\mathbf{d}_k^-$ are shown in (4.16) and (4.17).

$$\begin{aligned} \frac{\partial J_4}{\partial \mathbf{d}_t^+} &= -\sum_{i=1}^{K^+} \frac{1}{\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^b} \cdot \left(\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^{b-1} \cdot \frac{\partial \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)}{\partial \mathbf{d}_t^+} \right) \\ &\quad + \alpha \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} (\mathbf{D}^+ \mathbf{a}_{ij}^+)^T \mathbf{x}_{ij} \cdot a_{ijt}^+ \mathbf{x}_{ij} \end{aligned} \tag{4.16}$$

$$\frac{\partial J_4}{\partial \mathbf{d}_k^-} = -\sum_{i=1}^{K^+} \frac{1}{\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^b} \cdot \left(\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^{b-1} \cdot \frac{\partial \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)}{\partial \mathbf{d}_k^-} \right) + \rho \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} -2p_{ijk}\mathbf{q}_{ij} \tag{4.17}$$

The next step is taking the partial derivative of the hybrid detector in (4.7) with respect to $\mathbf{d}_t^+$ and $\mathbf{d}_k^-$ shown as Eq. (4.18) and (4.19) respectively:

$$\begin{aligned} \frac{\partial \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)}{\partial \mathbf{d}_t^+} &= \exp\left(-\beta \frac{\|\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}\|^2}{\|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^2}\right) \frac{\partial \left(-\beta \frac{\|\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}\|^2}{\|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^2}\right)}{\partial \mathbf{d}_t^+} \\ &= \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+) \frac{2\beta a_{ijt}^+ (\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij})}{\|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^2} \\ &= \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+) \frac{2\beta a_{ijt}^+ \mathbf{r}_{ij}}{\|\mathbf{q}_{ij}\|^2} \end{aligned} \tag{4.18}$$

$$\begin{aligned} &\frac{\partial \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)}{\partial \mathbf{d}_k^-} = \exp\left(-\beta \frac{\|\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}\|^2}{\|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^2}\right) \frac{\partial \left(-\beta \frac{\|\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}\|^2}{\|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^2}\right)}{\mathbf{d}_k^-} \\ = \;& \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+) \cdot \\ & \frac{2\beta a_{ijk}^- (\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}) \|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^2 - 2\beta p_{ijk} \|\mathbf{x}_{ij} - \mathbf{D}\mathbf{a}_{ij}\|^2 (\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij})}{\|\mathbf{x}_{ij} - \mathbf{D}^-\mathbf{p}_{ij}\|^4} \\ = \;& \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+) \frac{2\beta a_{ijk}^- \mathbf{r}_{ij} \|\mathbf{q}_{ij}\|^2 - 2\beta p_{ijk} \|\mathbf{r}_{ij}\|^2 \mathbf{q}_{ij}}{\|\mathbf{q}_{ij}\|^4} \end{aligned} \tag{4.19}$$

Substituting the gradient of hybrid detector with respect to $\mathbf{d}_t^+$ and $\mathbf{d}_k^-$ in Eq. (4.18) and (4.19) to Eq. (4.16) and (4.17), respectively, we can get the resultant gradient of the

objective function (4.11) over $\mathbf{d}_t^+$ and $\mathbf{d}_k^-$:

$$\begin{aligned}\triangle \mathbf{d}_t^+ &= -\sum_{i=1}^{K^+} \frac{1}{\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^b} \left(\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^b \cdot \frac{2\beta a_{ijt}^+ \mathbf{r}_{ij}}{\|\mathbf{q}_{ij}\|^2} \right) \\ &\quad +\alpha \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} (\mathbf{D}^+ \mathbf{a}_{ij}^+)^T \mathbf{x}_{ij} \cdot a_{ijt}^+ \mathbf{x}_{ij} \end{aligned} \tag{4.20}$$

$$\begin{aligned}\triangle \mathbf{d}_k^- &= -\sum_{i=1}^{K^+} \frac{1}{\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^b} \left(\sum_{j=1}^{N_i} \Lambda(\mathbf{x}_{ij}, \mathbf{D}|\mathbf{B}_i^+)^b \cdot 2\beta \frac{a_{ijk}^- \mathbf{r}_{ij} \|\mathbf{q}_{ij}\|^2 - p_{ijk} \|\mathbf{r}_{ij}\|^2 \mathbf{q}_{ij}}{\|\mathbf{q}_{ij}\|^4} \right) \\ &\quad -\rho \sum_{i=K^++1}^{K^++K^-} \sum_{j=1}^{N_i} 2 p_{ijk} \mathbf{q}_{ij} \end{aligned} \tag{4.21}$$

4.2.2 Optimization for Sparse Representation

The optimization of sparse representation can be viewed as a l_1 regularized least squares problem, also known as the lasso problem [116, 117, 123], denoted as $\mathcal{L}$. The lasso problem is shown in Eq. (4.8), where given concept (or dictionary) set $\mathbf{D}$ and preset sparsity level λ, $\mathbf{a}^*$ is the optimal sparse representation of the input data $\mathbf{x}$. Here we adopt the iterative shrinkage-thresholding algorithm (ISTA) [113, 114] for solving the sparse codes $\mathbf{a}$.

The gradient of (4.8) with respect to $\mathbf{a}$ without considering the l_1 penalty term is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = -\mathbf{D}^T (\mathbf{x} - \mathbf{D}\mathbf{a}) . \tag{4.22}$$

Then $\mathbf{a}$ at q^{th} iteration can be updated using gradient descent shown in (4.23):

$$\mathbf{a}^q = \mathbf{a}^{q-1} - \delta \frac{\partial \mathcal{L}}{\partial \mathbf{a}}, \tag{4.23}$$

followed by a soft-thresholding step:

$$\mathbf{a}^* = S_\lambda \left(\mathbf{a}^q \right), \tag{4.24}$$

where S_λ: $\mathbb{R}^n \rightarrow \mathbb{R}^n$ is the *soft-thresholding* operator defined by

$$S_\lambda \left(\mathbf{a}[k] \right) = sign(\mathbf{a}[k]) \max(|\mathbf{a}[k]| - \lambda, 0), k = 1, \cdots, n \tag{4.25}$$

Following a similar proof to that in [115], when the step length δ satisfies (4.26), the update of $\mathbf{a}$ using a gradient descent method with step length η monotonically decreases the value of the objective function, where $Eig_{max}(\mathbf{D^T D})$ denotes the maximum eigenvalue of $\mathbf{D}^T\mathbf{D}$. For simplicity, δ was set to $Eig_{max}\left(\mathbf{D}^T\mathbf{D}\right)$ for all input data:

$$\delta \in \left(0, \frac{1}{Eig_{max}\left(\mathbf{D}^T\mathbf{D}\right)} \right) \tag{4.26}$$

Finally the resultant update equation for the sparse representation of instance $\mathbf{x}$ given concept set $\mathbf{D}$ is:

$$\mathbf{a}^* = S_\lambda \left(\mathbf{a}^q + \frac{1}{Eig_{max}\left(\mathbf{D}^{-T}\mathbf{D}\right)} \left(\mathbf{D}^T(\mathbf{x} - \mathbf{D}\mathbf{a}^q) \right) \right) \tag{4.27}$$

4.3 Algorithm and Initialization

The algorithm for MI-HE is shown in Alg. 4

Algorithm 4 MI-HE algorithm

Initialize $\mathbf{D}^0$, $iter = 0$
repeat
  for $t = 1, \cdots, T$ **do**
    Solve $\mathbf{a}_{ij}$, $\mathbf{p}_{ij}$ according to (4.27), $\forall i \in \{1, \cdots, K\}, j \in \{1, \cdots, N_i\}$
    Update $\mathbf{d}_t^+$ using gradient descent according to (4.20)
    $\mathbf{d}_t^+ \leftarrow \frac{1}{\|\mathbf{d}_t^-\|_2}\mathbf{d}_t^+$
  end for
  for $k = T+1, \cdots, T+M$ **do**
    Solve $\mathbf{a}_{ij}$, $\mathbf{p}_{ij}$ according to (4.27), $\forall i \in \{1, \cdots, K\}, j \in \{1, \cdots, N_i\}$
    Update $\mathbf{d}_k^-$ using gradient descent according to (4.21)
    $\mathbf{d}_k^- \leftarrow \frac{1}{\|\mathbf{d}_k^-\|_2}\mathbf{d}_k^-$
  end for
  $iter \leftarrow iter + 1$
until Stopping criterion reached
return $\mathbf{D}$

The stopping criterion is either the change in objective function value is smaller than a preset threshold or the preset maximum number of iterations is reached.

The initialization of target concepts in $\mathbf{D}$ is conducted by computing the mean of T random subsets drawn from the union of all positive training bags. K-means [124] or VCA (vertex component analysis [109]) was applied to the union of all negative bags and the M cluster centers (or vertices) were set as the initial background concepts.

4.4 Classification using Estimated Concepts

Once the concept set $\mathbf{D}$ has been estimated using MI-HE, target detection on test data can be performed using the hybrid detector shown in (4.28), the ratio of the reconstruction errors given the target and non-target concepts vs. the non-target concepts only :

$$\Lambda_{HSD}(\mathbf{x}, \mathbf{D}) = \frac{(\mathbf{x} - \mathbf{D}^{-}\mathbf{p})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{D}^{-}\mathbf{p})}{(\mathbf{x} - \mathbf{D}\mathbf{a})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{D}\mathbf{a})}. \tag{4.28}$$

This provides a confidence that the i^{th} instance belongs in the target class (*i.e.*, if the numerator has a large error and the denominator has a small error, then the target concepts are needed in the reconstruction of the i^{th} instance).

Chapter 5

Experimental Results

The discriminative target concept learning algorithm MI-HE was applied to simulated hyperspectral data, the MUUFL Gulfport hyperspectral data set, ballistocardiogram data, hyperspectral data from NEON (National Ecological Observatory Network), and compared with our previously proposed algorithms and the state-of-the-art MIL algorithms. The area under the receiver operating characteristic (ROC) curve (AUC) or the normalized AUC (NAUC) was adopted as performance analysis metrics.

5.1 Hyperspectral Target Detection from Simulated Hyperspectral Data

MI-HE was first tested with simulated hyperspectral data. Assume given a set of candidate endmembers $\mathbf{D} = [\mathbf{D}^+, \mathbf{D}^-]$, each data point $\mathbf{x}$ is generated by the linear combination of the subset elements of $\mathbf{D}$ and some corresponding proportion values $\mathbf{a}$ following Eq. (4.5) (for target points) or Eq. (4.6) (for non-target points), respectively. The constituent

endmembers and the number of endmembers for each data point were drawn randomly. The corresponding background proportion values were generated by drawing from a Dirichlet distribution. The $\boldsymbol{\alpha}_{mean}$ parameter in the Dirichlet distribution was the expected mean value for the generating process with different level of variance controlled by the magnitude of $\boldsymbol{\alpha}_{mean}$. For a more precise description of how the simulated data was generated, pseudo-code describing the generation method is shown in Alg. 5 and 6.

Alg. 6 describes how each simulated data point is generated according to its bag-level label and instance-level label following the model in (4.5) and (4.6). Alg. 5 describes one method to generate simulated MIL data bags given the target signature set ($\mathbf{D}^+ \in \mathbb{R}^d \times T$), background signature set ($\mathbf{D}^- \in \mathbb{R}^{d \times M}$), number of positive bags (K^+), number of negative bags (K^-), number of points in each bag (N_i), the number of target points in each positive bag (N_{tar}), the minimum number of background endmembers per data point (N_b), the mean target proportion value ($\boldsymbol{\alpha}_{t\_mean}$), and a parameter to control proportion variance (σ). The code produces the following outputs: $\mathbf{x}$: a synthetic data vector, $\mathbf{X}$: full synthetic data matrix, $\mathbf{L}$: binary bag-level labels; and $\mathbf{l}$: binary instance-level labels.

5.1.1 Simulated Data with Incomplete Background Knowledge

As discussed in section 4, eFUMI combines all positive bags as one big positive bag and all negative bags as one big negative bag and learns target concept from the big positive bag that is different from the negative bag. So if the negative bags maintain incomplete knowledge of the background and only by distinguishing the commonly shared concept by positive bags the target concept can be correctly estimated, eFUMI will suffer from this kind of MIL problem. However, MI-HE which maintains bag structure will be able to estimate the target.

Algorithm 5 *Pseudo Code for Generating Synthetic Data as Bags*

Input: $\mathbf{D}^+, \mathbf{D}^-, K^+, K^-, N_i, N_{tar}, N_b, \boldsymbol{\alpha}_{t\_mean}, \sigma$
Output: $\mathbf{X} = \cup \mathbf{x}_{ij}, \mathbf{L}, \mathbf{l}$

for $i \leftarrow 1$ to K^+ **do**
    $\mathbf{L}(i) = 1$
    for $j \leftarrow 1$ to N_{tar} **do**
        $\mathbf{l}(i,j) = 1$
        $\mathbf{x}_{i,j} \leftarrow$ Alg. 6 given parameters set $\{\mathbf{D}^+, \mathbf{D}^-, \mathbf{L}(i), \mathbf{l}(i,j), N_b, \boldsymbol{\alpha}_{t\_mean}$ and $\sigma\}$
    end for
    for $j \leftarrow N_{tar} + 1$ to N_i **do**
        $\mathbf{l}(i,j) = 0$
        $\mathbf{x}_{i,j} \leftarrow$ Alg. 6 given parameters set $\{\mathbf{D}^+, \mathbf{D}^-, \mathbf{L}(i), \mathbf{l}(i,j), N_b, \boldsymbol{\alpha}_{t\_mean}$ and $\sigma\}$
    end for
end for
for $i \leftarrow K^+ + 1$ to $K^+ + K^-$ **do**
    $\mathbf{L}(i) = 0$
    for $j \leftarrow 1$ to N_i **do**
        $\mathbf{l}(i,j) = 0$
        $\mathbf{x}_{i,j} \leftarrow$ Alg. 6 given parameters set $\{\mathbf{D}^+, \mathbf{D}^-, \mathbf{L}(i), \mathbf{l}(i,j), N_b, \boldsymbol{\alpha}_{t\_mean}$ and $\sigma\}$
    end for
end for

Algorithm 6 *Pseudo Code for Generating Linearly Mixed Data Point Given Bag-level and Point-level Label*

Input: $\mathbf{D}^+, \mathbf{D}^-, L_i, l_{ij}, N_b, \boldsymbol{\alpha}_{t\_mean}, \sigma$
Output: $\mathbf{x}$

if $L_i \& l_{ij}$ **then**
    Draw an random integer t between $[1, T]$
    Randomly select t dictionary elements (denoted as $\mathbf{D}_t^+$) from $\mathbf{D}^+$
    Draw an random integer m between $[N_b, M]$
    if $m == 0$ **then**
        $\boldsymbol{\alpha}_{mean} = \boldsymbol{\alpha}_{t\_mean}$
    else
        $\boldsymbol{\alpha}_{mean} = \sigma \cdot [\boldsymbol{\alpha}_{t\_mean}, \frac{1-\boldsymbol{\alpha}_{t\_mean}}{m} \times \mathbf{1}_{1\times m}]$
    end if
    Randomly select m dictionary elements (denoted as $\mathbf{D}_m^-$) from $\mathbf{D}^-$
    $\boldsymbol{\alpha} \leftarrow$ sample $t + m$ random values from Dirichlet Distribution given parameter $\boldsymbol{\alpha}_{mean}$
    Generate point $\mathbf{x}$ following the Linear Mixing Model (4.5) using $\mathbf{a}$ and $[\mathbf{D}_t^+, \mathbf{D}_m^-]$
else
    Uniformly draw integer m between $[max(1, N_b), M]$
    $\boldsymbol{\alpha}_{mean} = \sigma \cdot \mathbf{1}_{1\times m}$
    Randomly select m dictionary elements (denoted as $\mathbf{D}_m^-$) from $\mathbf{D}^-$
    $\boldsymbol{a} \leftarrow$ sample m random values from Dirichlet Distribution given parameter $\boldsymbol{\alpha}_{mean}$
    Generate point $\mathbf{x}$ following the Linear Mixing Model (4.6) using $\mathbf{a}$ and $\mathbf{D}_m^-$
end if

Given this hypothesis, simulated data was generated from four spectra selected from the ASTER spectral library [125]. Specifically, the Red Slate, Verde Antique, Phyllite and Pyroxenite spectra from the rock class with 211 bands and wavelengths ranging from 0.4μm to 2.5μm (as shown in Fig. 5.1) were used as endmembers to generate hyperspectral data. Red Slate was labeled as the target endmember.

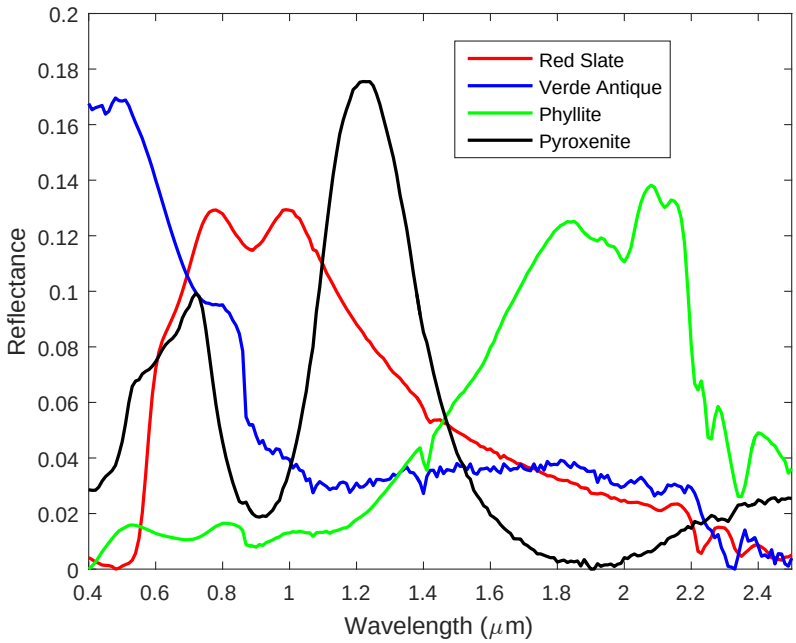


Figure 5.1: Signatures from ASTER library used to generate simulated data with incomplete background knowledge

Four sets of highly-mixed noisy data with varied mean target proportion value ($\boldsymbol{\alpha}_{t\_mean}$) were generated according to Alg. 5 and 6. Specifically, this synthetic data has $K^+ = 15$ positive and $K^- = 5$ negative bags with each bag has $N_i = 500$ points. If it is a positively labeled bag, there are $N_{tar} = 200$ highly-mixed target points with mixture level $N_b = 1$ containing mean target (Red Slate) proportion $\boldsymbol{\alpha}_{t\_mean}$. The parameter $\boldsymbol{\alpha}_{t\_mean}$ that controls the mean target proportion value was set to 0.1, 0.3, 0.5 and 0.7 respectively to vary the

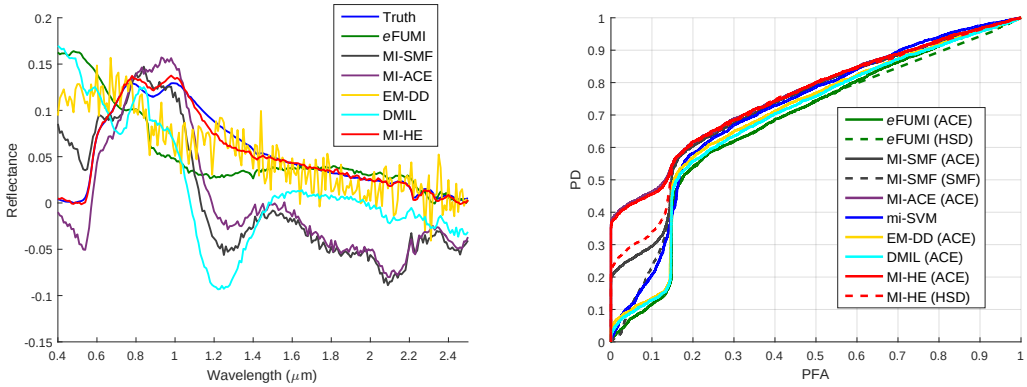
Table 5.1: List of Constituent Endmembers for Synthetic Data with Incomplete Background Knowledge

Bag Number	Bag Label	Target Endmember	Background Endmember
1-5	+	Red Slate	Verde Antique, Phyllite, Pyroxenite
6-10	+	Red Slate	Phyllite, Pyroxenite
11-15	+	Red Slate	Pyroxenite
16-20	−	N/A	Phyllite, Pyroxenite

level of target presence from weak to high. The scale factor that controls the variance of the Dirichlet distribution is set to $\sigma = 2$. Gaussian white noise was added so that signal-to-noise ratio of the data was set to $20dB$. To highlight the ability of MI-HE to leverage individual bag-level labels, we use different subsets of background endmembers to build synthetic data as shown in Tab. 5.1. Tab. 5.1 shows that the negatively labeled bags only contain 2 negative endmembers and there exists one confusing background endmember in the first 5 positive bags which is Verde Antique. However, only the target endmember Red Slate was placed in all 15 positive bags and the learner is expected to distinguish Red Slate correctly by checking what is commonly shared over positive bags. So it is expected that the proposed MI-HE will be able to learn the target signature correctly and eFUMI will confuse both Red Slate and Verde Antique as target signatures since Verde Antique is missing in the training negative bags (and eFUMI does not preserve bag-level labels).

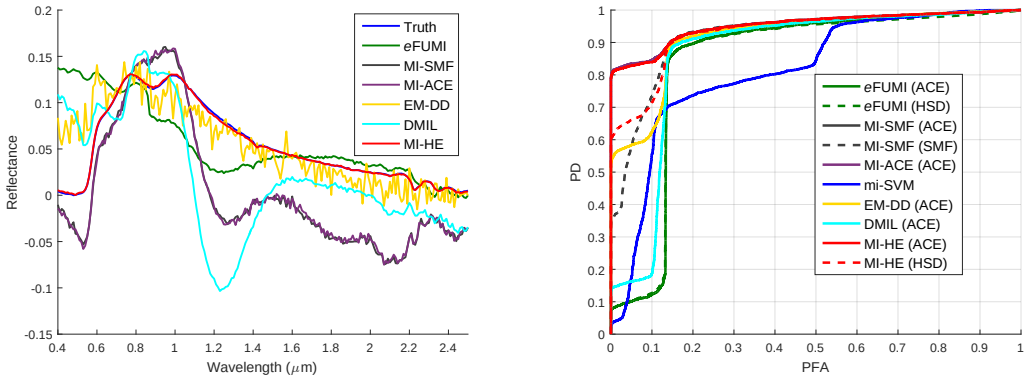
The parameter settings of MI-HE for this experiment are $T = 1, M = 9, \rho = 0.8, p = 5, \beta = 5$ and $\lambda = 1 \times 10^{-3}$. MI-HE was compared to our previously proposed algorithm eFUMI, MI-SMF, and MI-ACE and several state-of-the-art MIL algorithms DMIL, EM-DD and mi-SVM. The mi-SVM algorithm was added to these experiments to include a comparison MIL approach that does not rely on estimating a target signature.

Fig. 5.2(a) - 5.5(a) show the estimated target signature from simulated hyperspectral data with different levels of mean target presence from 0.1 to 0.7. Those figures clearly



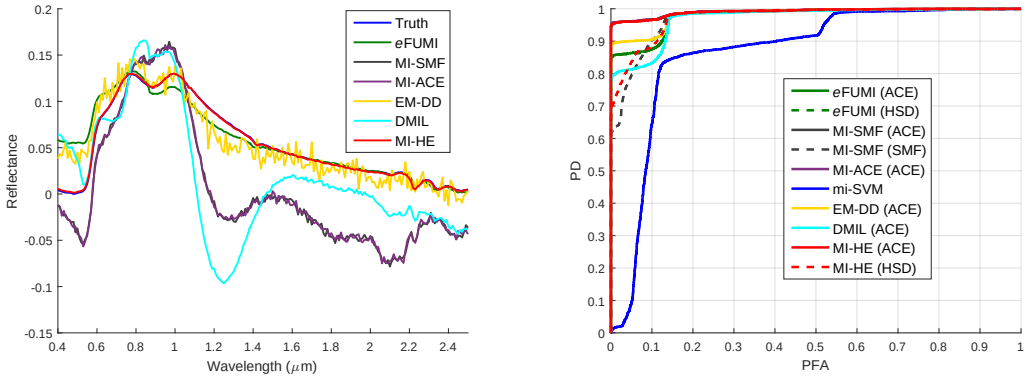
(a) Estimated target signatures (b) ROC curves

Figure 5.2: MI-HE and comparisons on synthetic data with incomplete background knowledge, $\boldsymbol{\alpha}_{t\_mean} = 0.1$. MI-SMF and MI-ACE are not expected to recover the true signature.



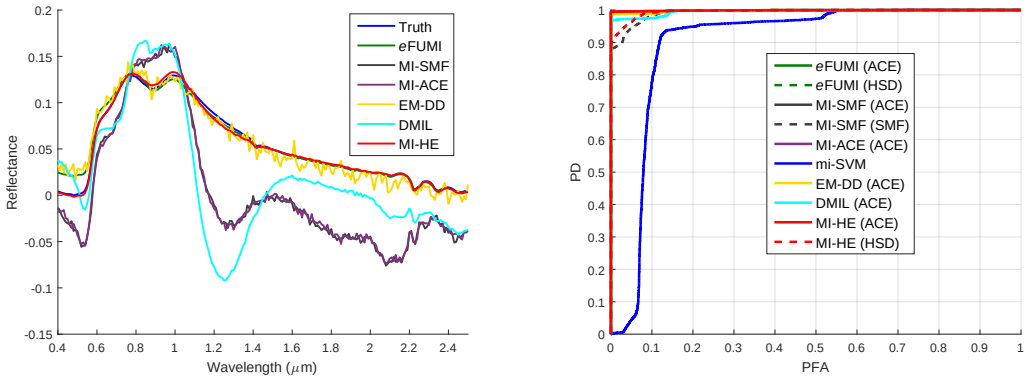
(a) Estimated target signatures (b) ROC curves

Figure 5.3: MI-HE and comparisons on synthetic data with incomplete background knowledge, $\boldsymbol{\alpha}_{t\_mean} = 0.3$. MI-SMF and MI-ACE are not expected to recover the true signature.



(a) Estimated target signatures (b) ROC curves

Figure 5.4: MI-HE and comparisons on synthetic data with incomplete background knowledge, $\boldsymbol{\alpha}_{t\_mean} = 0.5$. MI-SMF and MI-ACE are not expected to recover the true signature.



(a) Estimated target signatures (b) ROC curves

Figure 5.5: MI-HE and comparisons on synthetic data with incomplete background knowledge, $\boldsymbol{\alpha}_{t\_mean} = 0.7$. MI-SMF and MI-ACE are not expected to recover the true signature

show that the proposed MI-HE is able to correctly distinguish Red Slate as target concept from the incomplete background knowledge. Also, the other comparison algorithms can also estimate a target concept close to the ground truth Red Slate spectrum. However, eFUMI is always confused with the other non-target endmember, Verde Antique, that exists in some positive bags but is excluded from the background bags.

For simulated detection analysis, estimated target concepts from the training data were then applied to the test data generated separately following the same generating procedure as training data. The detection was performed using the HSD or ACE detection statistic. For MI-HE and eFUMI, both methods were applied since those two algorithms can come out a set of background concept from training simultaneously; for MI-SMF, both SMF and ACE were applied since MI-SMF's objective is maximizing the multiple instance spectral matched filter; for the rest multiple instance target concept learning algorithms, MI-ACE, EM-DD, DMIL, only ACE was applied. For the testing procedure of mi-SVM, a regular SVM testing process was performed using LIBSVM [126], and the decision values (signed distances to the separating hyperplane) of test data determined from trained SVM model were taken as the confidence values. For the signature based detectors, the background data mean and covariance were estimated from the negative instances of the training data.

For quantitative evaluation, Fig. 5.2(b) - 5.5(b) show the receiver operating characteristic (ROC) curves using estimated target signature, where it can be seen that the eFUMI is confused with the testing Verde Antique data at very low PFA rate. Tab. 5.2 shows the AUCs of proposed MI-HE and comparison algorithms. The results reported are the median results over five runs of the algorithm on the same data. From Tab. 5.2, it can be seen that the proposed MI-HE and our previously proposed discriminative multiple instance concept learning algorithms MI-ACE and MI-SMF achieved the best performance on detection us-

Table 5.2: Detection Statistics (AUCs) for Simulated Hyperspectral Data with Incomplete Background Knowledge, Bold for the Best, Underline for the Second Best

Algorithm	$\boldsymbol{\alpha}_{t\_mean}$			
	0.1	0.3	0.5	0.7
MI-HE (HSD)	0.743	<u>0.931</u>	0.975	0.995
MI-HE (ACE)	<u>0.763</u>	**0.952**	**0.992**	**0.999**
eFUMI (ACE)	0.675	0.845	0.978	<u>0.998</u>
eFUMI (HSD)	0.671	0.564	0.978	<u>0.998</u>
MI-SMF (SMF)	0.719	0.923	0.972	0.993
MI-SMF (ACE)	0.735	**0.952**	**0.992**	**0.999**
MI-ACE (ACE)	**0.764**	**0.952**	**0.992**	**0.999**
mi-SVM	0.715	0.815	0.866	0.900
EM-DD (ACE)	0.695	0.918	<u>0.983</u>	<u>0.998</u>
DMIL (ACE)	0.687	0.865	0.971	0.996

ing ACE detector. The reason that MI-HE's detection performance using HSD detector is little worse is that HSD relies on knowing the complete background concept to properly represent each non-target testing data, the missing non-target concept (Verde Antique) makes the non-target testing data containing Verde Antique maintain a large reconstruction error, and thus large detection statistics.

5.1.2 Simulated Data with Multiple Target Concepts

For this experiment, we select another rock endmember, Quartz Conglomerate, from ASTER spectral library as the second target concept. Three sets of highly-mixed noisy data with varied mean target proportion value ($\boldsymbol{\alpha}_{t\_mean}$) from [0.1, 0.1] to [0.3, 0.3] were generated according to Alg. 5 and 6. Specifically, this synthetic data has $K^+ = 5$ positive bags containing both target concept (Red Slate and Quartz Conglomerate) and non-target concept (Verde Antique, Phyllite, Pyroxenite); $K^- = 5$ negative bags containing only the back-

ground concept (Verde Antique, Phyllite, Pyroxenite); each bag has $N_i = 500$ points. If it is a positively labeled bag, there are $N_{tar} = 200$ highly-mixed target points with mixture level $N_b = 1$. The parameter $\boldsymbol{\alpha}_{t\_mean}$ that controls the mean target proportion value was set to [0.1, 0.1], [0.2, 0.2] and [0.3, 0.3], respectively to vary the level of target presence from weak to high. The scaling factor that controls the variance of the Dirichlet distribution is set to $\sigma = 2$. Gaussian white noise was added so that signal-to-noise ratio of the data was set to $20dB$. Tab. 5.3 shows the constituent endmembers for each bags. It is expected that the proposed MI-HE is able to learn multiple target concept at one time.

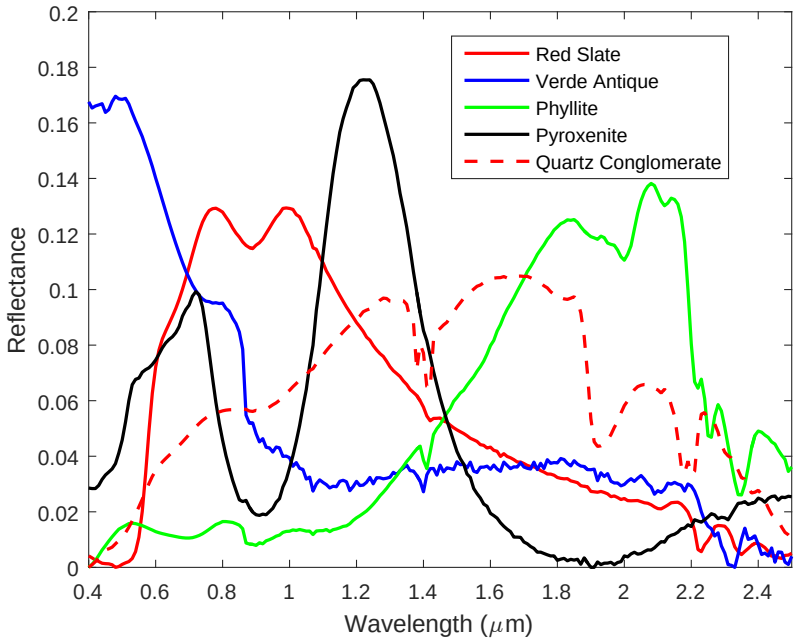
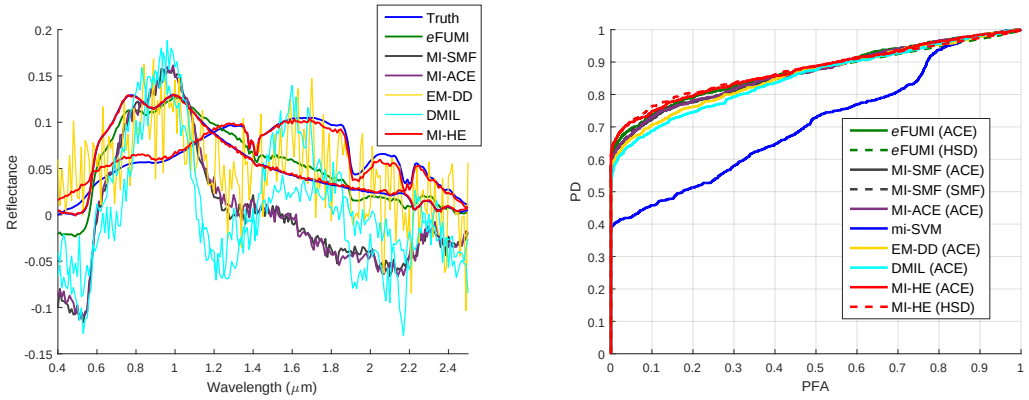


Figure 5.6: Signatures from ASTER library used to generate simulated data with multiple target concepts

The parameter settings of MI-HE for this experiment are $T = 2, M = 9, \rho = 0.8, p = 5, \beta = 5$ and $\lambda = 1 \times 10^{-3}$. Fig. 5.7(a) - 5.9(a) show the estimated target concept by proposed MI-HE and comparisons, where we can see that the proposed MI-HE is able

Table 5.3: List of Constituent Endmembers for Synthetic Data with Multiple Target Concepts

Bag Number	Bag Label	Target Endmembers	Background Endmembers
1-5	+	Red Slate, Quartz Conglomerate	Verde Antique, Phyllite, Pyroxenite
6-10	−	N/A	Verde Antique Phyllite, Pyroxenite



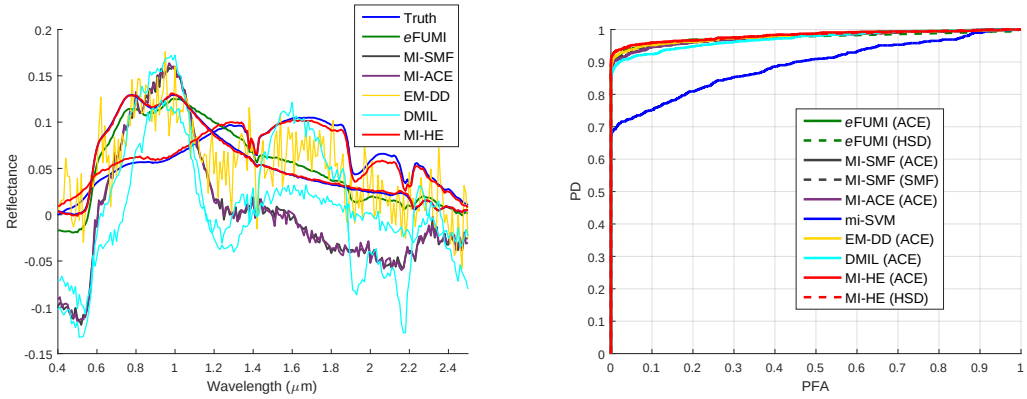
(a) Estimated target signatures

(b) ROC curves

Figure 5.7: MI-HE and comparisons on synthetic data with multiple target concepts, $\boldsymbol{\alpha}_{t\_mean} = [0.1, 0.1]$. Not all comparisons algorithms are expected to recover true target signatures.

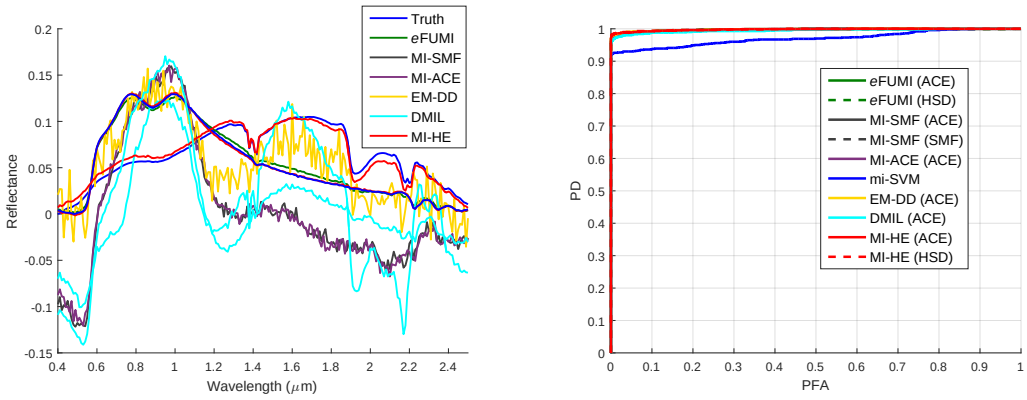
to accurately estimate multiple target concepts simultaneously. Compared with MI-HE, although DMIL is also a multiple concept learning algorithm, target concept as estimated by DMIL is noisy, and not a representative prototype of the target class. The remaining comparison algorithms are single target concept learning which are always confused by the multiple target concept problem.

For quantitative evaluation, Fig. 5.7(b) - 5.9(b) show the ROCs using estimated target signature, and Tab. 5.4 shows the AUCs of proposed MI-HE and comparison algorithms. The results reported are the median results over five runs of the algorithm on the same data. For ACE detector using multiple estimated target concept, the maximum detection



(a) Estimated target signatures (b) ROC curves

Figure 5.8: MI-HE and comparisons on synthetic data with multiple target concepts, $\boldsymbol{\alpha}_{t\_mean} = [0.2, 0.2]$. Not all comparisons algorithms are expected to recover true target signatures.



(a) Estimated target signatures (b) ROC curves

Figure 5.9: MI-HE and comparisons on synthetic data with multiple target concepts, $\boldsymbol{\alpha}_{t\_mean} = [0.3, 0.3]$. Not all comparisons algorithms are expected to recover true target signatures.

Table 5.4: Detection Statistics (AUCs) for Simulated Hyperspectral Data with Multiple Target Concepts, Bold for the Best, Underline for the Second Best

Algorithm	$\boldsymbol{\alpha}_{t\_mean}$		
	[0.1, 01]	[0.2, 0.2]	[0.3, 0.3]
MI-HE (HSD)	**0.875**	**0.982**	**0.998**
MI-HE (ACE)	**0.875**	**0.982**	**0.998**
eFUMI (ACE)	<u>0.872</u>	<u>0.980</u>	**0.998**
eFUMI (HSD)	0.865	0.976	<u>0.997</u>
MI-SMF (SMF)	0.866	0.977	<u>0.997</u>
MI-SMF (ACE)	0.865	0.976	<u>0.997</u>
MI-ACE (ACE)	0.866	0.976	<u>0.997</u>
mi-SVM	0.711	0.890	0.970
EM-DD (ACE)	0.858	0.979	**0.998**
DMIL (ACE)	0.850	0.971	0.994

statistics across all estimated target concept was selected for each testing data. From Tab. 5.4, it can be seen that the proposed MI-HE outperforms all the comparison single MI concept learning algorithms as well as the multiple MI concept learning algorithm DMIL.

5.1.3 Analysis of MI-HE Parameter Settings on Simulated Data

In order to provide deeper insights into the sensitivity of MI-HE performance relative to variations in input parameters, we tested MI-HE on simulated hyperspectral data across a range of parameter values. Specifically, the varying parameters and ranges examined are:

$M \in [1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21]$;

$\beta \in [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100]$;

$\lambda \in [1\times10^{-4}, 2\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}, 2\times10^{-3}, 5\times10^{-3}, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1]$;

$p \in [-10, -5, -2, -1, 1 \times 10^{-10}, 1, 2, 5, 10, 20, 50, 100]$.

Because the generalized mean function is not continuous at 0, 1×10^{-10} was chosen instead.

Red Slate was again selected as target endmember and the other three, Verde Antique, Phyllite and Pyroxenite were used as non-target endmembers as shown in Fig. 5.1. Tab. 5.5 shows the bags labeling and constituent endmembers. The synthetic data has $K^+ = 5$ positive and $K^- = 5$ negative bags with each bag containing $N_i = 100$ points. If it is a positively labeled bag, there are $N_{tar} = 50$ highly-mixed target points with mixture level $N_b = 1$ containing mean target (Red Slate) proportion $\boldsymbol{\alpha}_{t\_mean} = 0.1$. The scale factor that controls the variance of the Dirichlet distribution is set to $\sigma = 2$. Gaussian white noise was added so that the signal-to-noise ratio of the data was set to $20dB$. The reference parameters for MI-HE were set to $T = 1, M = 7, \rho = 0.8, p = 5, \beta = 5$ and $\lambda = 1 \times 10^{-3}$.

Fig. 5.10 shows the detection performance of MI-HE for the sensitivity analysis on the simulated data described above. Several interesting inferences can be drawn from Fig. 5.10 regarding how MI-HE responds to its parameter settings. For the setting of M, it can be seen from Fig. 5.10(a) that MI-HE performs consistently given the number of background concepts greater or equal to 3, which is the true number of background concepts of the simulated data. Since MI-HE adopts a linear mixing model instead of convex mixing for the data mixture, the conclusion is MI-HE is not sensitive to the setting of M given M is set not smaller than the necessary constituent background concepts of input data.

For the setting of β, Fig. 5.10(b) shows MI-HE performs well with β in the range [1, 10] on this data. Since β is a scaling factor for the hybrid detector, $\Pr(\mathbf{x}_{ij}|\mathbf{D}, \mathbf{B}_i^+) = \exp\left(-\beta \frac{\|\mathbf{x}_{ij} - \mathbf{D}\boldsymbol{a}_{ij}^+\|^2}{\|\mathbf{x}_{ij} - \mathbf{D}^-\boldsymbol{p}_{ij}\|^2}\right)$, the scaling effect controls the gradient of the objective function (4.11). Fig. 5.11 shows the plot of $exp(-\beta)$ with varying β values, which shows that the range $\beta \in$ [1, 10] provides a moderate gradient for the exponential function. So it can be concluded that MI-HE requires β to be set to a suitable range, *e.g.*, [1, 10].

The setting of sparsity level, λ, results in the step length of soft shrinkage. This value

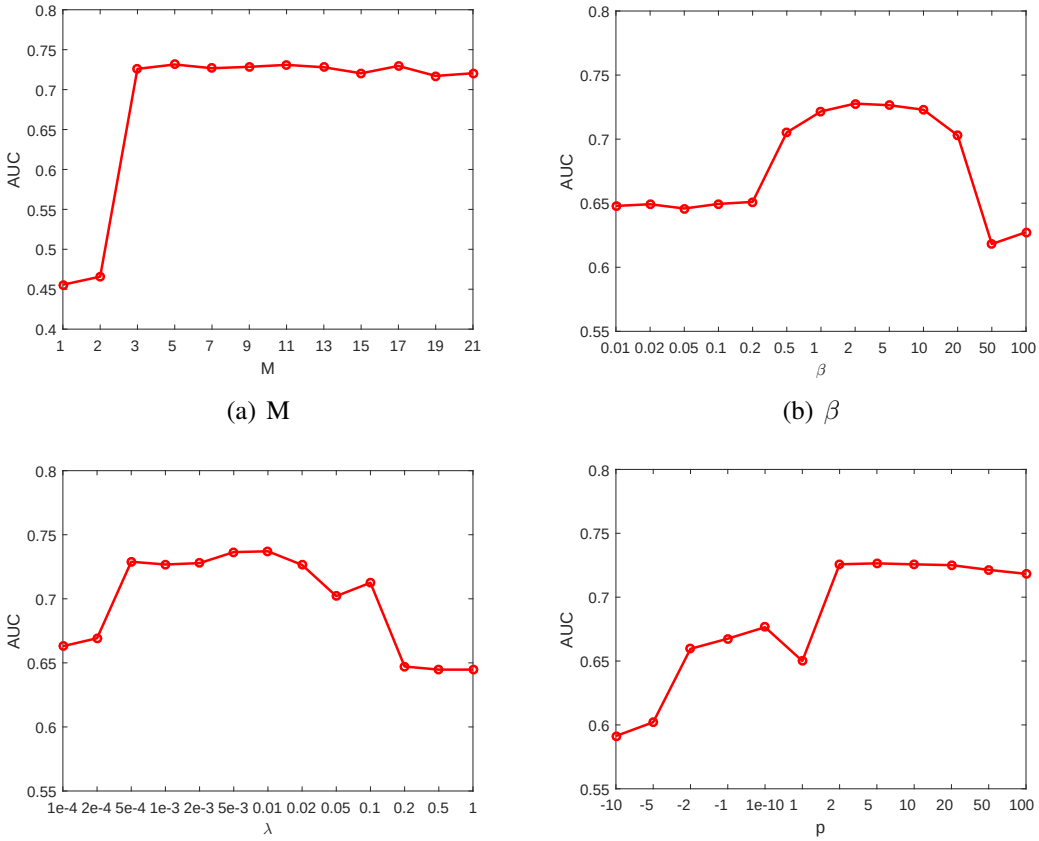
Table 5.5: List of Constituent Endmembers for Synthetic Data for Parameter Sensitivity Testing

Bag Number	Bag Label	Target Endmember	Background Endmembers
1-5	$+$	Red Slate	Verde Antique, Phyllite, Pyroxenite
6-10	$-$	N/A	Verde Antique Phyllite, Pyroxenite

should be related with the magnitude of the input data. For the simulated data tested here, the proportion values were generated from the Dirichlet distribution within range [0, 1]. As shown by Fig. 5.10(c), the appropriate range of λ for this simulated data is $[5 \times 10^{-4}, 0.02]$ which is reasonable. However, in general, a prior knowledge is needed for setting specific λ for a real dataset. Currently we set λ to approximately the $1/1000$ of the l_2-norm mean of the training data.

The parameter p is related to the effects of the generalized mean model from minimum ($p \to -\infty$), or mean ($p \in (0, 1]$) to maximum ($p \to +\infty$). Since the proposed model aims to predict the true positive instance from positive bags and assumes the "soft maximum" operation for this generalized mean model, it is expected that the model will work well with p greater than 1. Fig. 5.10(d) verifies this hypothesis showing that the algorithm works well for p great than 1. For all experiments shown in this paper, the parameter p was set to 5 and observed to work well.

Although there are several parameters for MI-HE, these parameters come from the models assumed to underlie MI-HE. From this sensitivity analysis of MI-HE with different parameter settings, it can be concluded that there is a general range of model stability for each parameter. Moreover, the above analysis provides a intuitive understanding and heuristic approach for setting the parameters of MI-HE to stable ranges.



(a) M (b) β

(c) λ (d) p

Figure 5.10: Detection statistics (AUCs) of MI-HE plots with different parameter settings

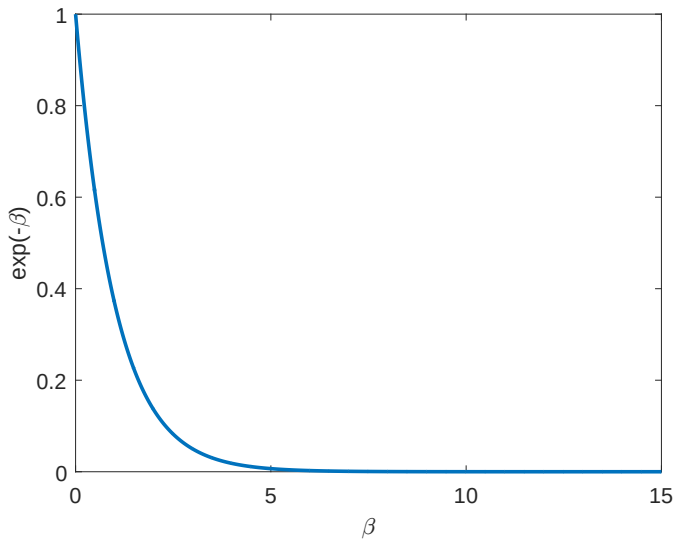


Figure 5.11: Plot of exponential function $exp(-\beta)$

5.2 Hyperspectral Target Detection from Real Hyperspectral Data

For experiments on real hyperspectral target detection data, the MUUFL Gulfport hyperspectral data set collected over the University of Southern Mississippi-Gulfpark Campus was used. This data set contains 325×337 pixels with 72 spectral bands corresponding to wavelengths from $367.7nm$ to $1043.4nm$ at a $9.5 - 9.6nm$ spectral sampling interval with spatial resolution 1 pixel$/m^2$ [51]. The first four and last four bands were removed due to sensor noise. Two sets of this data (Gulfport Campus Flight 1 and Gulfport Campus Flight 3) were selected as cross-validated training and testing data for these two data sets have the same altitude and spatial resolution. Throughout the scene, there are 64 man-made targets in which 57 were considered in this experiment which are cloth panels of four different colors: Brown (15 examples), Dark Green (15 examples), Faux Vineyard Green (FVGr)

(12 examples) and Pea Green (15 examples). The spatial location of the targets are shown as scattered points over an RGB image of the scene in Fig. 5.12. Some of the targets are in the open ground and some are occluded by the live oak trees. Moreover, the targets also vary in size, for each target type, there are targets that are $0.25m^2$, $1m^2$ and $9m^2$ in area, respectively, resulting a very challenging, highly mixed sup-pixel target detection problem.

5.2.1 MUUFL Gulfport Hyperspectral Data, Individual Target Type Detection

For this part of the experiments, each individual target type was treated as a target class, respectively. For example, when "Brown" is selected as target class, a 5×5 rectangular region corresponding to each of the 15 ground truth locations denoted by GPS was grouped into a positive bag to account for the drift coming from GPS. This size was chosen based on the accuracy of the GPS device used to record the ground truth locations. The remaining area that does not contain a brown target was grouped into a big negative bag. This constructs the detection problem for "Brown" target. Similarly, there are 15, 12, 15 positive labeled bags for Dark Green, Faux Vineyard Green and Pea Green, respectively. The parameter settings of MI-HE for this experiment are $T = 1, M = 9, \rho = 0.3, p = 5, \beta = 1$ and $\lambda = 5 \times 10^{-3}$.

MI-HE and comparison algorithms were evaluated on this data using the Normalized Area Under the receiver operating characteristic Curve (NAUC) in which the area was normalized out to a false alarm rate (FAR) of 1×10^{-3} false alarms$/m^2$ [127]. MI-HE was compared to the eFUMI, MI-SMF, MI-ACE, mi-SVM, EM-DD, DMIL algorithms. Target concepts were estimated on the training flight and then used to perform detection on the test flight using the HSD or ACE detection statistic. For MI-HE and eFUMI, both meth-

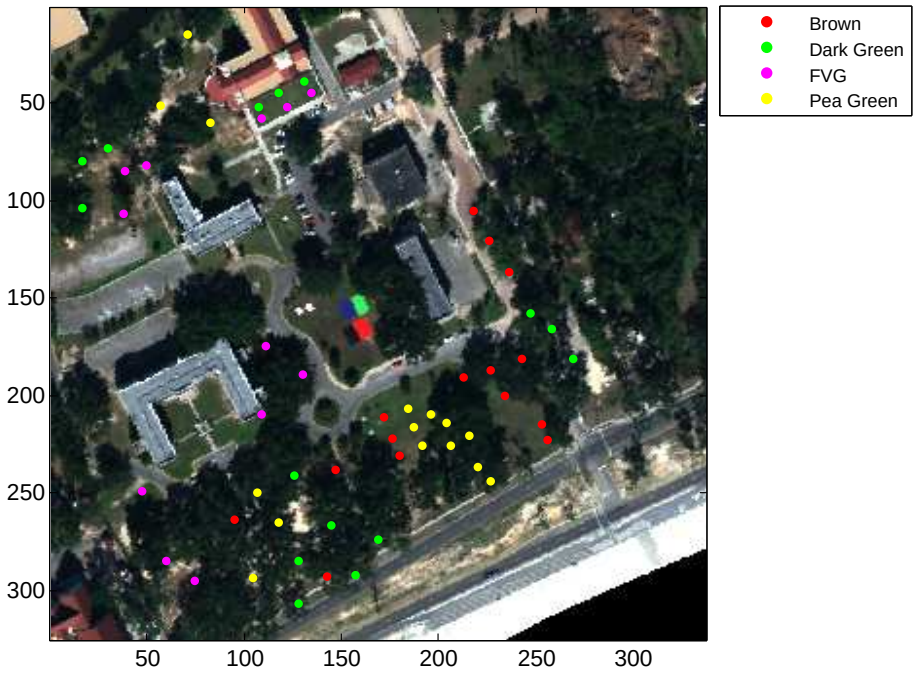
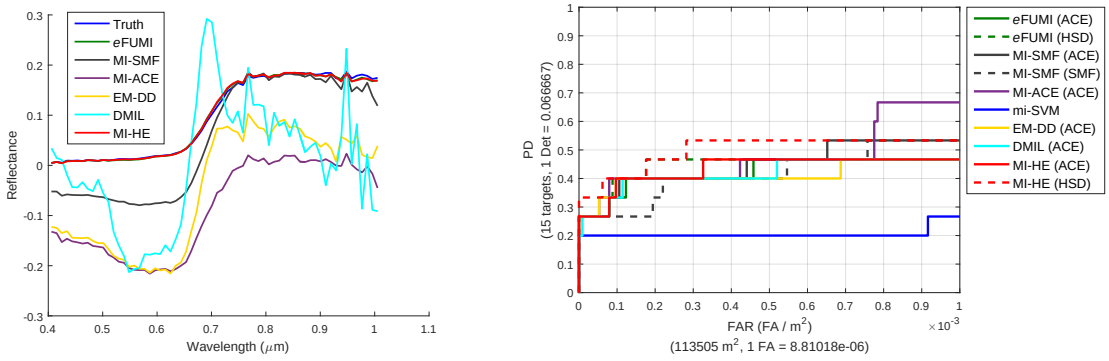


Figure 5.12: MUUFL Gulfport data set RGB image and the 57 target locations

ods were applied; for MI-SMF, both SMF and ACE were applied; for the other multiple concept learning algorithm, only ACE was applied; for mi-SVM, a regular SVM testing process was performed using LIBSVM and the decision values (signed distances to the separating hyperplane) of test data determined from the trained SVM model were taken as the confidence values. During detection on the test data, the background mean and covariance were estimated from the negative instances of the training data. The results reported are the median results over five runs of the algorithm on the same data.

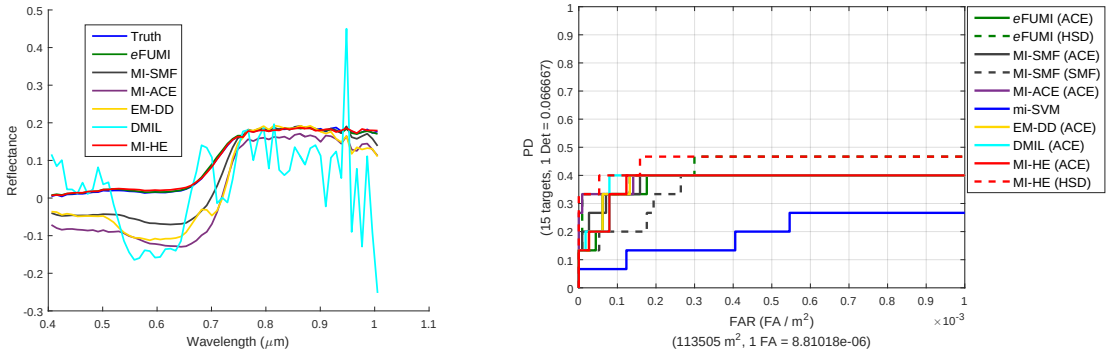
Fig. 5.13(a) - 5.20(a) show the estimated target concept by proposed MI-HE and comparisons for four types of target and two flights, respectively. We can see that the proposed MI-HE is able to recover the target concept quite close to ground truth spectra manually selected from the scene. Fig. 5.13(b) - 5.20(b) show the detection ROCs given target spectra estimated on one flight data and cross validated on another flight data. Tab. 5.6 shows the



(a) Estimated target signatures

(b) ROC curves

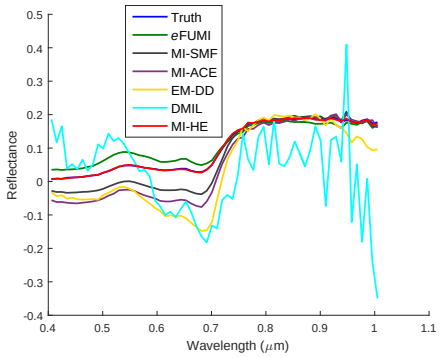
Figure 5.13: MI-HE and comparisons on Gulfport data Brown, training flight 1 testing flight 3



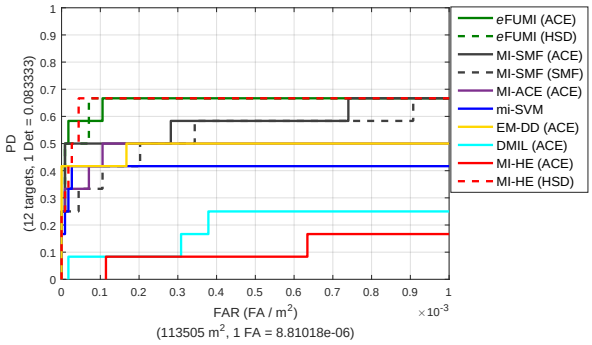
(a) Estimated target signatures

(b) ROC curves

Figure 5.14: MI-HE and comparisons on Gulfport data Dark Green, training flight 1 testing flight 3

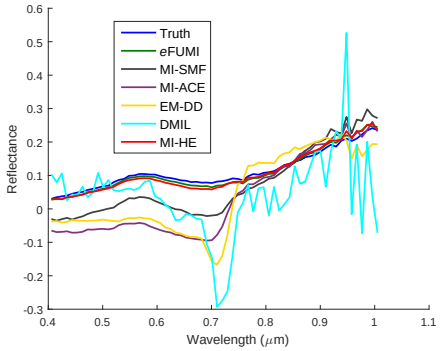


(a) Estimated target signatures

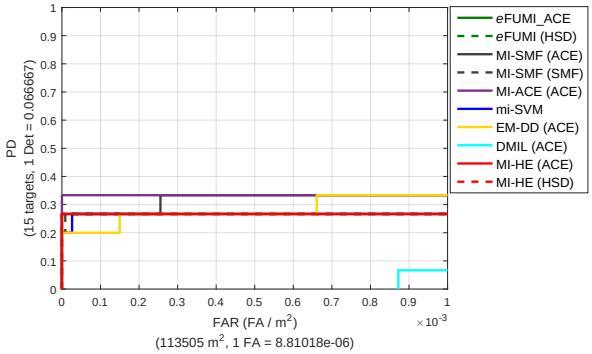


(b) ROC curves

Figure 5.15: MI-HE and comparisons on Gulfport data Faux Vineyard Green, training flight 1 testing flight 3

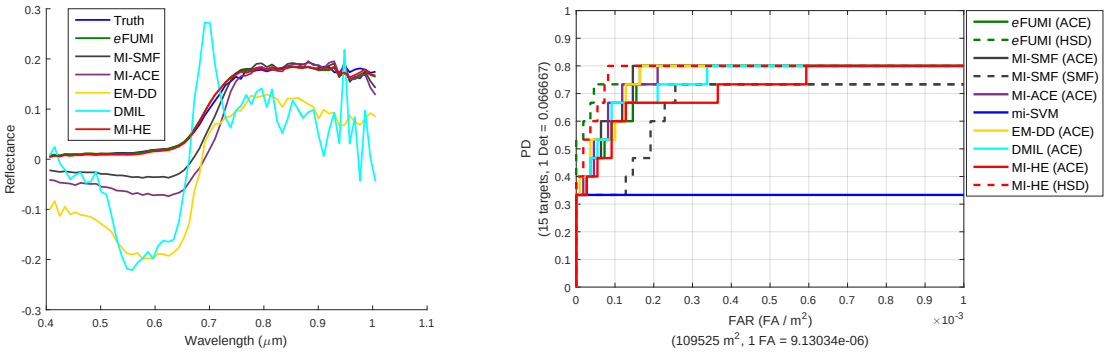


(a) Estimated target signatures



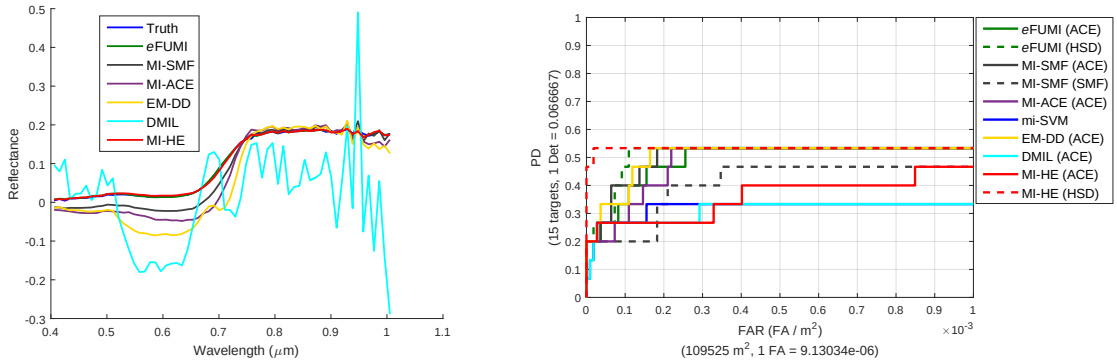
(b) ROC curves

Figure 5.16: MI-HE and comparisons on Gulfport data Pea Green, training flight 1 testing flight 3



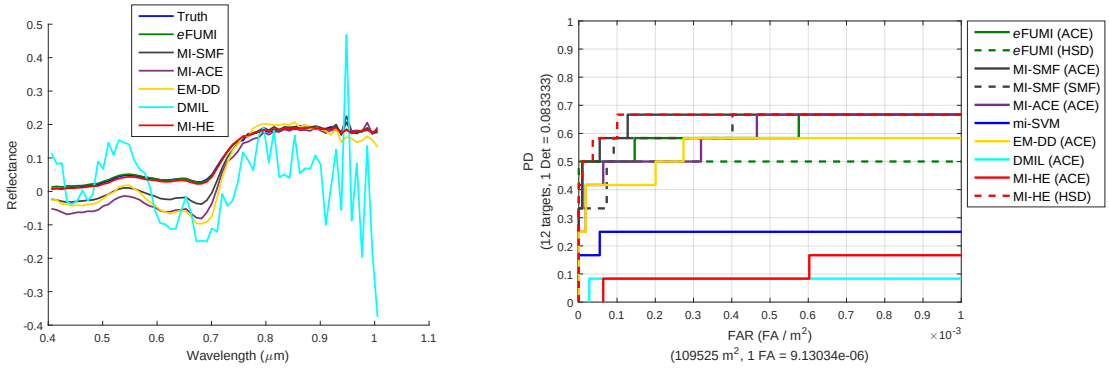
(a) Estimated target signatures (b) ROC curves

Figure 5.17: MI-HE and comparisons on Gulfport data Brown, training flight 3 testing flight 1



(a) Estimated target signatures (b) ROC curves

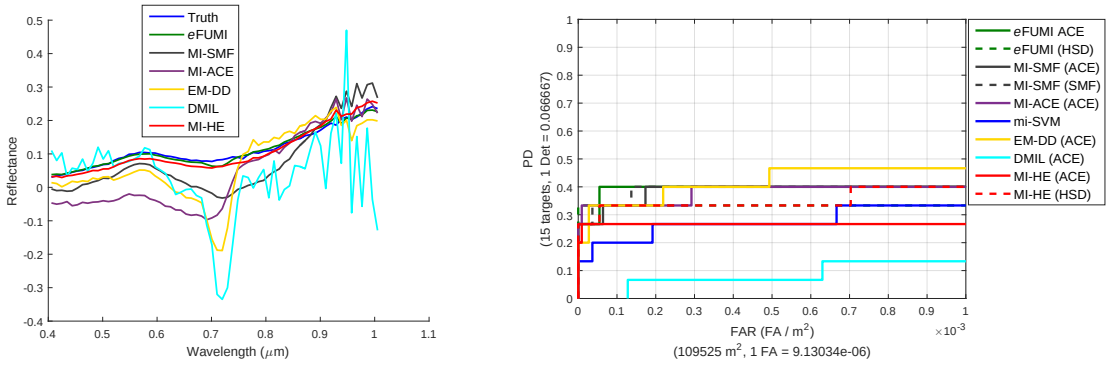
Figure 5.18: MI-HE and comparisons on Gulfport data Dark Green, training flight 3 testing flight 1



(a) Estimated target signatures

(b) ROC curves

Figure 5.19: MI-HE and comparisons on Gulfport data Faux Vineyard Green, training flight 3 testing flight 1



(a) Estimated target signatures

(b) ROC curves

Figure 5.20: MI-HE and comparisons on Gulfport data Pea Green, training flight 3 testing flight 1

NAUCs for MI-HE and comparison algorithms, where it can be seen the proposed MI-HE outperforms the comparisons for most of the target types.

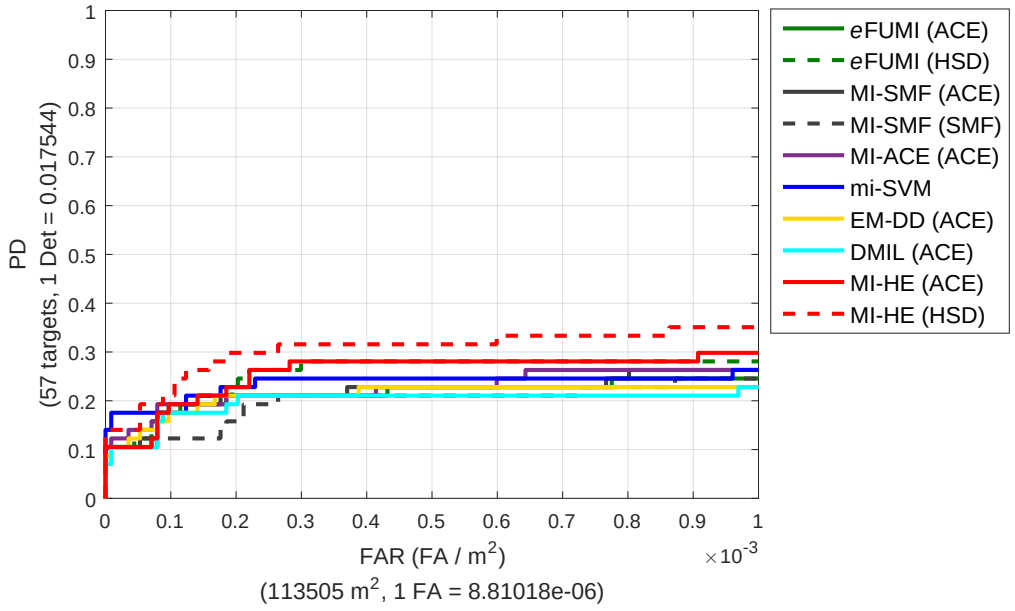
Table 5.6: Detection Statistics (NAUCs) for Gulfport Data with Individual Target Type, Bold for the Best, Underline for the Second Best

Alg.	Train on Flight 1; Test on Flight 3				Train on Flight 3; Test on Flight 1			
	Brown	Dark Gr.	Faux Vine Gr.	Pea Gr.	Brown	Dark Gr.	Faux Vine Gr.	Pea Gr.
MI-HE (HSD)	**0.499**	**0.453**	**0.655**	0.267	**0.781**	**0.532**	**0.655**	0.350
MI-HE (ACE)	0.433	0.379	0.104	0.267	0.710	0.360	0.111	0.266
***e*FUMI (ACE)**	0.423	0.377	<u>0.654</u>	0.267	0.754	0.491	0.605	<u>0.393</u>
***e*FUMI (HSD)**	0.444	<u>0.436</u>	0.653	0.267	0.727	<u>0.509</u>	0.500	0.333
MI-SMF (SMF)	0.419	0.354	0.533	0.266	0.657	0.405	<u>0.650</u>	0.384
MI-SMF (ACE)	0.448	0.382	0.579	<u>0.316</u>	<u>0.760</u>	0.501	0.613	0.388
MI-ACE(ACE)	<u>0.474</u>	0.390	0.485	**0.333**	<u>0.760</u>	0.483	0.593	0.380
mi-svm	0.206	0.195	0.412	0.265	0.333	0.319	0.245	0.274
EM-DD(ACE)	0.411	0.381	0.486	0.279	<u>0.760</u>	0.503	0.541	**0.416**
DMIL(ACE)	0.419	0.383	0.191	0.009	0.743	0.310	0.081	0.083

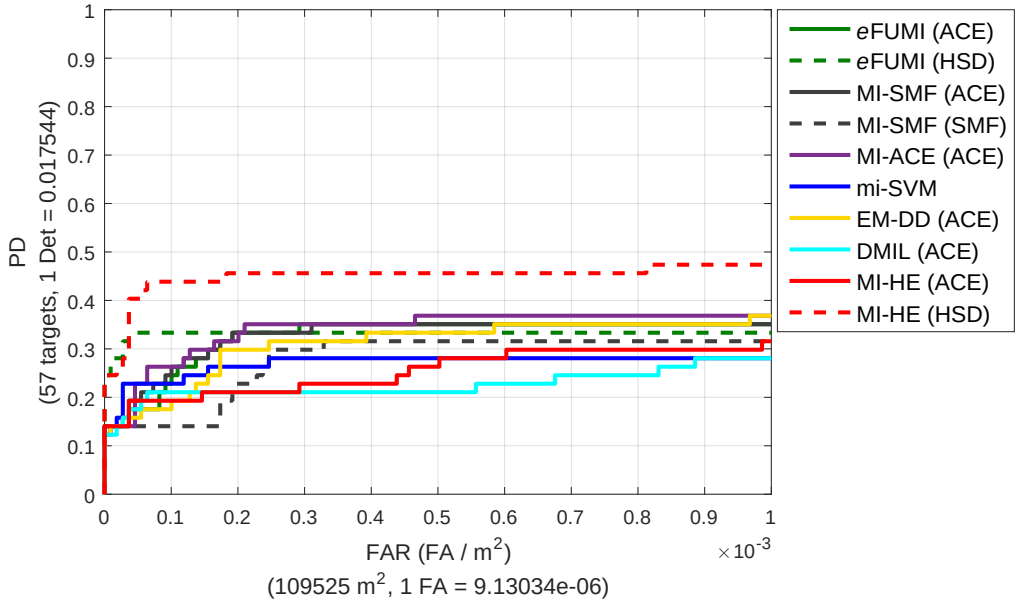
5.2.2 MUUFL Gulfport Hyperspectral Data, All Four Target Types Detection

For training and detection for the four target types together, the positive bags were generated by grouping each of the 5×5 regions denoted by the ground truth that it contains any of the four types of target. Thus, for each flight there are 57 targets and 57 positive bags were generated. The remaining area that does not contain any target was grouped into a big negative bag. The parameter settings of MI-HE for this experiment are $T = 9, M = 11, \rho = 0.3, p = 5, \beta = 1$ and $\lambda = 5 \times 10^{-3}$.

Fig. 5.21(a) - 5.21(b) show the detection ROCs given target spectra estimated on one flight data and cross validated on another flight data, which shows that given estimated multiple target concepts, the detection statistics by proposed MI-HE using HSD are significantly better than the comparison algorithms. Tab. 5.7 summarizes the NAUCs as a quantitative comparison.



(a) Train on flight 1, detect on flight 3



(b) Train on flight 3, detect on flight 1

Figure 5.21: ROCs of MI-HE and comparisons on Gulfport data, all types detection.

Table 5.7: Detection Statistics (NAUCs) for Gulfport Data with All Four Target Types, Bold for the Best, Underline for the Second Best

Alg.	Tr. Fl. 1; Te. Fl. 3	Tr. Fl. 3; Te. Fl. 1	Alg.	Tr. Fl. 1; Te. Fl. 3	Tr. Fl. 3; Te. Fl. 1
MI-HE (HSD)	**0.304**	**0.449**	**MI-SMF(ACE)**	0.219	0.327
M-IHE (ACE)	<u>0.257</u>	0.254	**MI-SMF(SMF)**	0.198	0.277
***e*FUMI (ACE)**	0.214	<u>0.325</u>	**mi-SVM**	0.235	0.269
***e*FUMI (HSD)**	0.256	0.331	**EM-DD(ACE)**	0.211	0.310
MI-ACE (ACE)	0.226	0.340	**DMIL(ACE)**	0.198	0.225

5.3 Beat-to-Beat Heart Rate Monitoring from Ballistocardiogram Data

In this work, we applied the proposed MI-HE to heartbeat detection and rate monitoring from ballistocardiogram signals. Given the hydraulic bed sensor system described in Sec. 1.2, in this study, we use the dataset collected from 40 subjects at the CERT at the University of Missouri. The data collection from human subjects has been approved by the Institutional Review Board (IRB) at the University of Missouri. To prepare for the data collection, each subject was asked to lie flat on their back for 10 minutes. The gender, age, weight and height of the subjects are listed in Table 5.8. There were 7 females and 33 males. Their ages, weights, heights and BMIs are 18 - 49, 48 - 127 (Kg), 156 - 190 (cm), 18.3 - 37.9, with the average as 29.2, 76.9 (Kg), 175.2 (cm) and 24.9, respectively. The BCG signal was sampled at 100Hz and filtered by a six order Butterworth band-pass filter with 3dB cutoff frequency at 0.4Hz and 10Hz to remove the respiratory component and high frequency noise.

Feature Extraction from Ballistocardiograms Time Series: The filtered time domain segments from received BCG signals were used as training features. Specifically, for each subject, we found the signal peaks for the entire BCG signal received by four transducers as candidate J-peak locations (possible heartbeat locations). For each peak, we extract a data segment (simply called instance) that is 91 samples long and centered

at the peak (corresponding to 0.91s signal, 45 samples before and after the peak) as the training feature for this peak location. The feature length (0.91s, at 100 Hz sampling rate) was determined empirically and found to be the typical length of a heartbeat pattern. Fig. 5.22 shows example filtered BCG signals collected by four transducers (blue plots) and the corresponding finger sensor ground truth information (black plots), where the green circles denote every peak location of the filtered BCG signal.

MIL Bags for Ballistocardiograms Time Series: In this study, we continue to investigate the idea of training "bags" to address label uncertainty as well as miss-collection of heartbeat signals in the BCG data. Specifically, each of the extracted sub-signals is treated as individual data points (or "instances") during training. Each positive labeled training bag (shown as red rectangles in Fig. 5.22) was formed by grouping the 12 instances across the four transducers (3 instances for each transducer) that are close in time to the ground-truth location marked by the finger sensor (shown as a red cross in Fig. 5.22). Similarly, one negative bag was formed by grouping instances from four transducers between two positive bags that were not included in any positive bag. Fig. 5.23 shows an example positive bag. In this figure, we can see that a positively labeled bag contains both true heartbeat patterns (shown in red) and non-heartbeat patterns (shown in dotted green). From Fig. 5.23 it can also be seen that the assumed heartbeat patterns tend to have more prominent J-peaks. The proposed MI-HE algorithm is expected to learn a set of discriminative subject-specific heartbeat concepts from training bags of this type. After learning the heartbeat concept, the HSD was applied for real-time heartbeat monitoring and heart rate estimation.

For each subject, the 10 minutes BCG signal was split into 5 minutes for training and 5 minutes for testing. The parameter settings of MI-HE for this experiment are $T = 3, M = 3, \rho = 0.3, p = 5, \beta = 1$ and $\lambda = 5 \times 10^{-3}$. Fig. 5.24 shows estimated heartbeat concepts

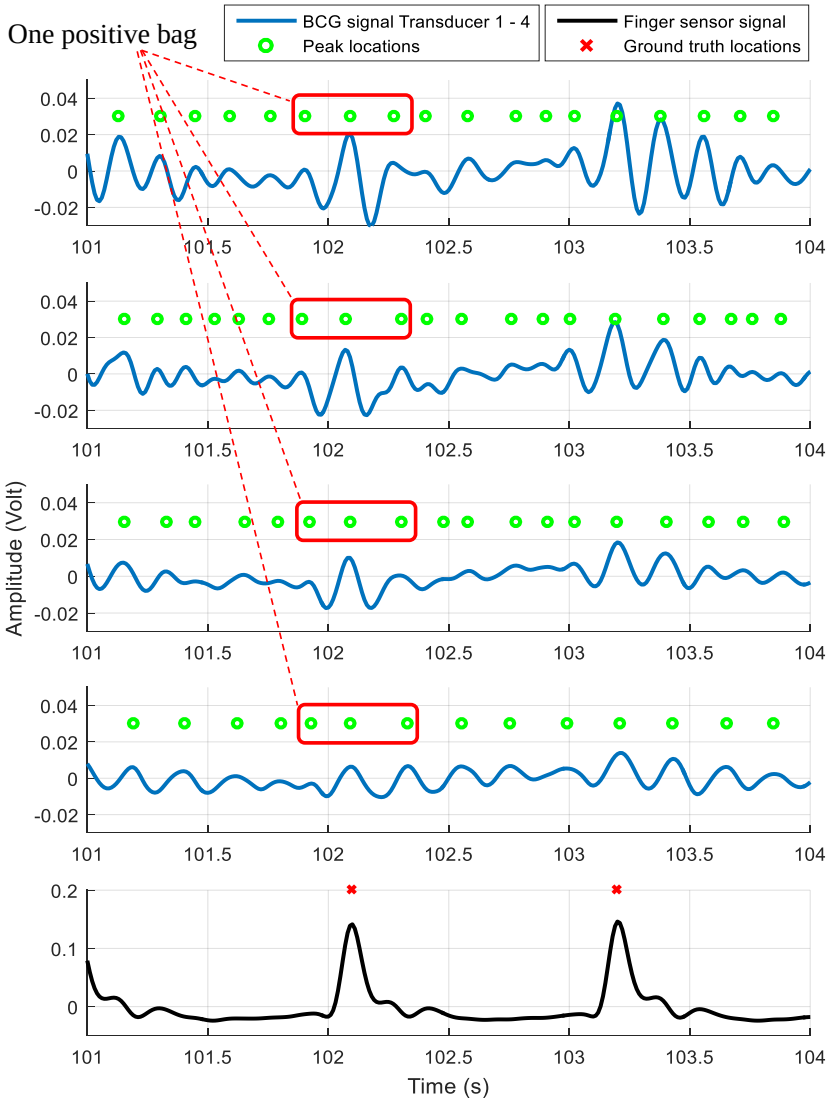


Figure 5.22: BCG signal of four transducers and ground truth plot.

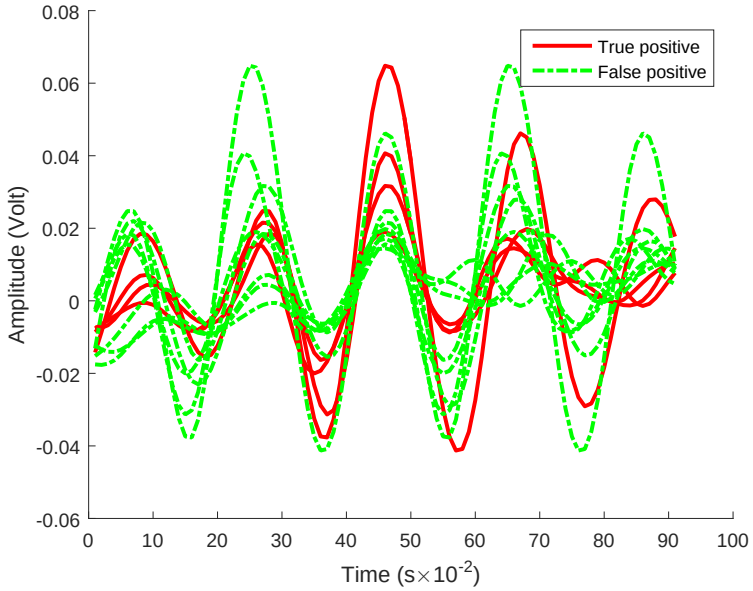


Figure 5.23: Plot of one positive bag.

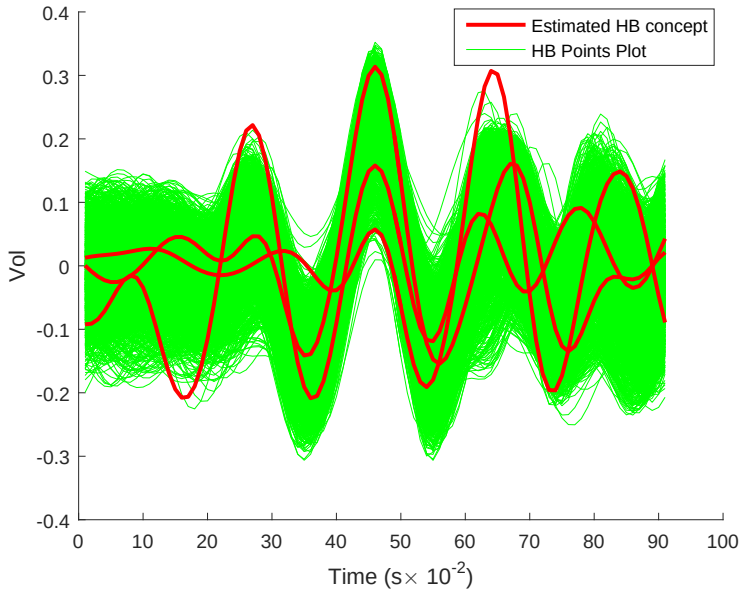


Figure 5.24: Estimated heartbeat concepts by MI-HE.

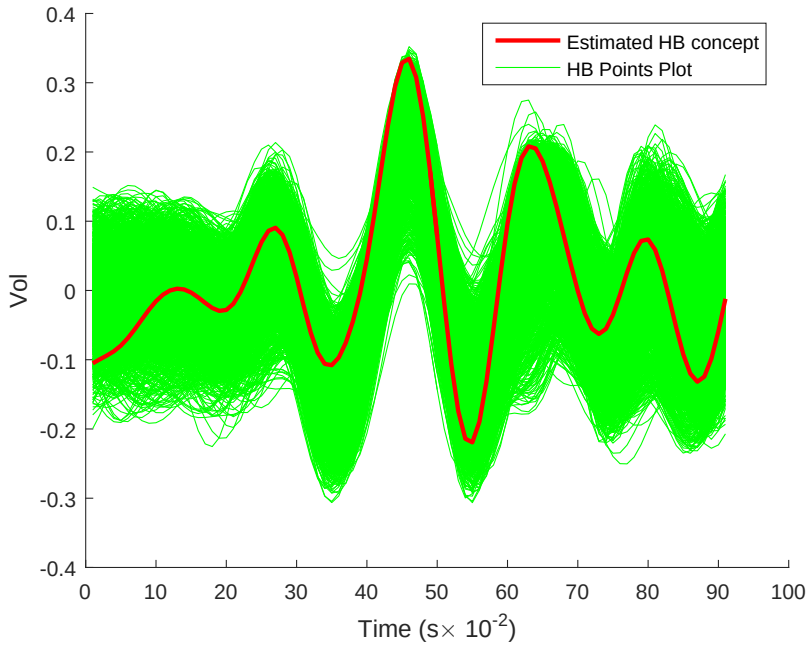


Figure 5.25: Estimated heartbeat concept by EM-DD.

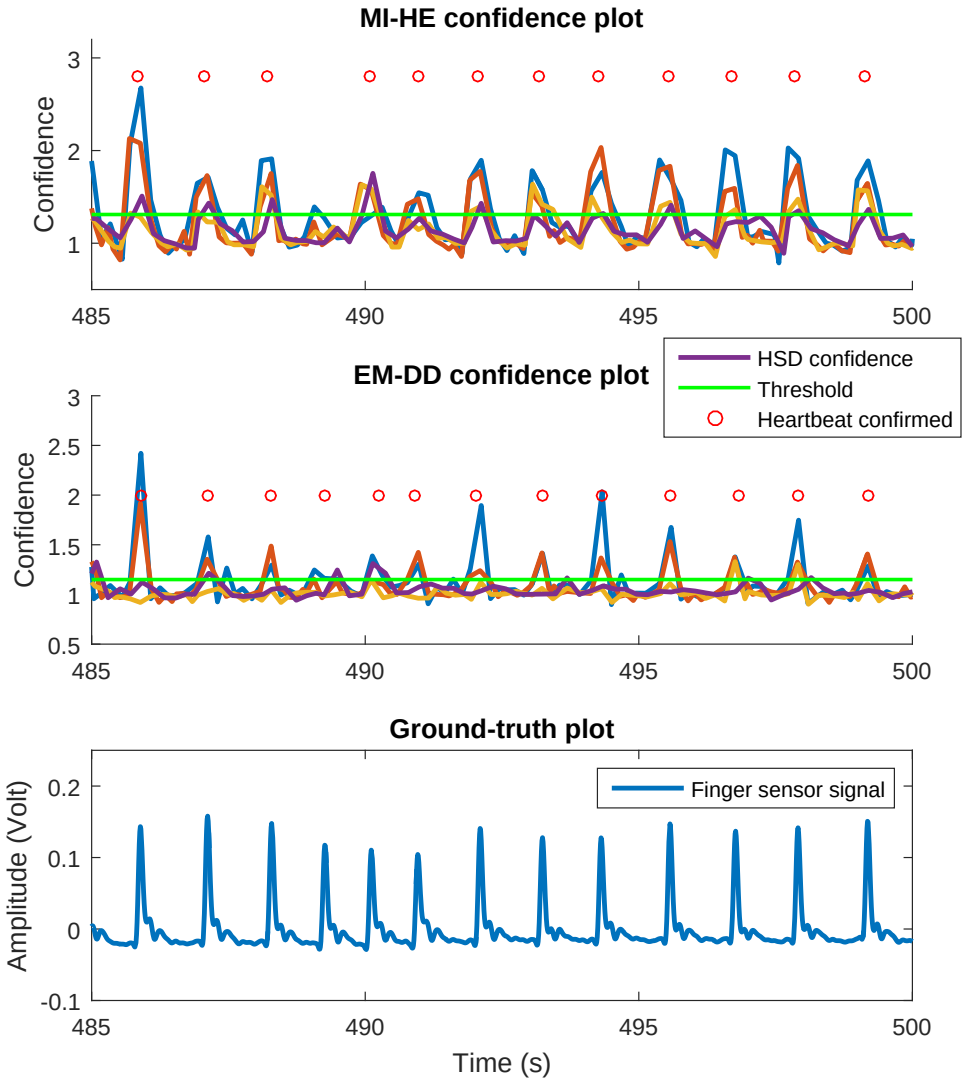


Figure 5.26: Confidence value and confirmed heartbeats.

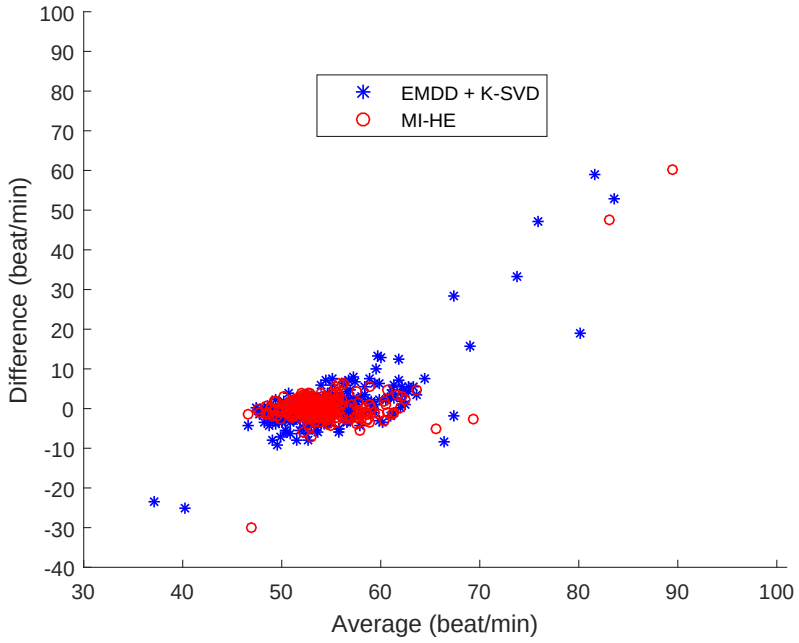
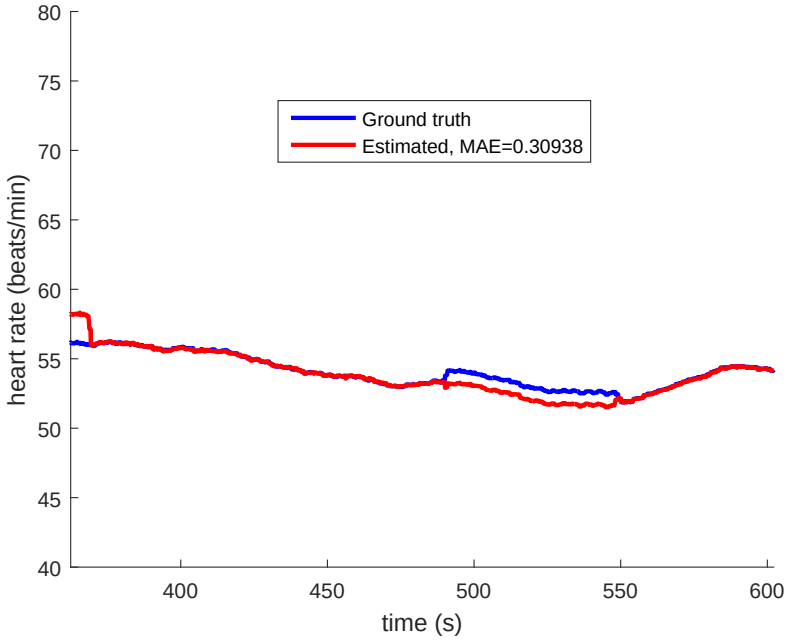


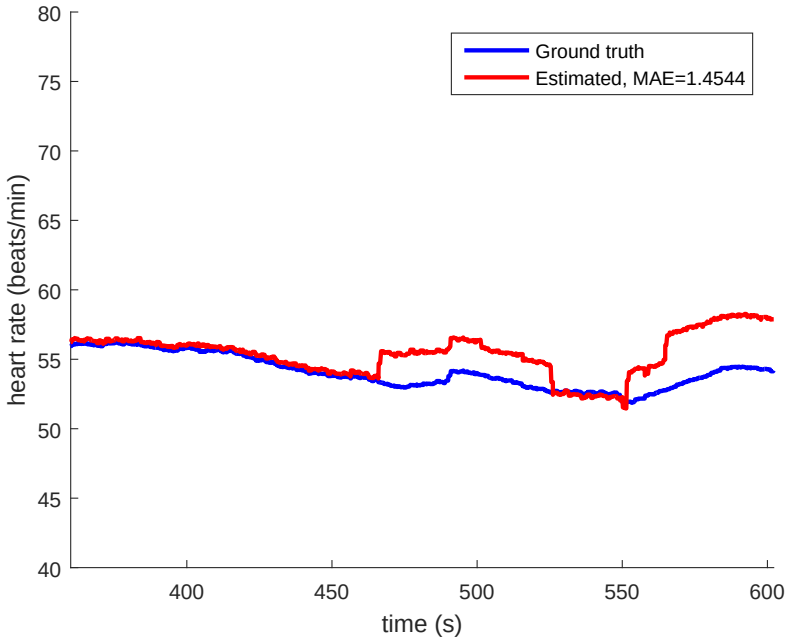
Figure 5.27: The Bland Altman plot comparison of MI-HE and EM-DD for Subject No. 10.

for subject No. 10 as an example, where we can see the heartbeat concept estimated by MI-HE maintains prominent J-peaks. For comparison, we applied EM-DD [78] which is a widely used multiple instance concept learning algorithm to the same data. From Fig. 5.24 and 5.25 we can clearly see that although both the heartbeat concepts estimated by MI-HE and EM-DD have prominent J-peaks, MI-HE is able to learn multiple concepts to account for the variability in heartbeat prototype during sleeping, which helps improve performance in heartbeat detection and rate estimation.

After learning heartbeat concepts, heartbeat detection on test data can be carried out. In the results shown in this paper, the HSD was applied to the test data to get a confidence value for each data point to be a true heartbeat signal. Since EM-DD only learns a single target concept, we applied K-SVD [128] which is a widely used unsupervised dictionary



(a) MI-HE



(b) EM-DD

Figure 5.28: Heart rate estimation for Subject No. 10. (a) MI-HE (b) EM-DD

learning algorithm to the negative labeled training bags to get a set of non-target concepts for EM-DD heartbeat detection using HSD. Fig. 5.26 shows the excerpt of HSD 4 - channel confidence value (485s - 500s) for subject No. 10 estimated by the heartbeat concepts of MI-HE (shown in Fig. 5.24) and EM-DD (shown in Fig. 5.25), respectively. In this procedure, a heartbeat (J-peak) is confirmed through a voting procedure requiring at least two confidence values within a neighborhood (25 samples) that are greater than a threshold (1.31) across all four transducers. The neighborhood and threshold values are determined via cross-validation on training data. From Fig. 5.26 we can see that the confidence peaks estimated by MI-HE match the peaks of the finger sensor signal very well. A missed detection was found at about 489s. For comparison, the confidence estimated by EM-DD is not as prominent as MI-HE and there are several false alarms found. Fig. 5.27 shows the beat to beat Bland Altman plot of MI-HE and EM-DD, where we can see the plot of MI-HE is more compact than that of EM-DD. For MI-HE, the outlier in the bottom left corresponding to the missed detection appeared around 489s in Fig. 5.26. The two outliers in the top right corresponding to the one false alarm found by MI-HE around 310s.

For heart rate estimation, the average of beat-to-beat heart rates over 1 minute is computed using a sliding window. Fig. 5.28(a) and Fig. 5.28(b) show the estimated heart rate on testing data for subject No. 10 by MI-HE and EM-DD, respectively. From Fig. 5.28(a) we can see that the drop in estimated heart rate around 489s comes from the missed detection of heartbeat in Fig 5.26, the rise at the beginning of estimated heart rate corresponding to one false alarm appeared in 310s. As a comparison, the rise in estimated heart rate in Fig. 5.28(b) comes from the several false alarms estimated by EM-DD.

A comprehensive study and comparison with more state-of-the-art BCG heart rate monitoring algorithms were conducted on the 40 subjects. The four algorithms used for the

comparison are window-peak-to-peak (WPPD) [129], clustering approach (CA) [53], energy (EN) [130], and Hilbert transform (HT) [55, 131]. The details of results are listed in Table 5.8. The mean error shown in Table 5.8 is the mean absolute error (MAE) between the estimation and ground truth at every 15s interval, where the bold numbers highlight the best and underlined numbers highlight the second best performance for each subject. Over 40 subjects, the mean error of proposed algorithm is 0.64 (beat/min), which is the best over the comparison algorithms. The mean error of EM-DD+K-SVD, EN, WPPD, CA and HT are 2.43, 3.02, 1.42, 1.49 and 0.83 (beat/min), respectively. For EN, there are 3 subjects that have aberrant estimation (with error greater than 10) and make the average result worse. If we ignore these three outliers, the mean error of EN is 2.04 (beat/min).

To examine if the performance of proposed algorithm is related to subject's age, weight, height, BMI, and magnitude of subject's heart rate, we compute the correlation coefficient between the results in Table 5.8 and subjects' age, weight, height, BMI and the average of ground truth heart rate over testing data (denoted as average GT). The sample Pearson correlation coefficient r defined in Eq. (5.1) measures the linear correlation between two variables x and y, where i is the index of the variables and $\bar{x}$ and $\bar{y}$ are the mean of two variables, respectively. Table 5.9 shows the corresponding correlation coefficient of the proposed algorithm and the other five comparison algorithms. For weight, height, and BMI, all six algorithms show no relation. For age, all algorithms consistently show negative relationship with error, this is mainly due to the increase in heartbeat variations with the decrease in age. However, compared with the other five algorithms, the proposed MI-HE and our previous DL-FUMI are more robust to age since MI-HE and DL-FUMI [72] are able to learn multiple concepts to account for heartbeat variability. Based on the results, we can tell that the performance is not related to age, weight, height, and BMI, which validates

Table 5.8: Errors of MI-HE and Comparisons for Heart Rate Monitoring from 40 Subjects, Bold for the Best, Underline for the Second Best

Subject	Age, sex (F/M), BMI weight (kg), height (cm)	Mean Error (beat/min)						
		MI-HE	DL-FUMI	EM-DD+K-SVD	EN	WPPD	CA	HT
1	43, M, 23.3, 74.8, 179	0.91±0.11	**0.19**±7.95×10^{-3}	0.96±0.82	0.22	1.26	1.76±0.09	0.79
2	21, M, 23.1, 74, 179	0.64±0.09	0.85±0.05	4.99±1.48	**0.58**	1.67	1.68±0.14	0.94
3	33, M, 27.2, 75, 166	0.27±0.07	0.22±0.20×10^{-3}	0.71±0.94	**0.17**	1.02	1.05±0.03	0.77
4	23, F, 21.1, 52, 157	1.55±0.06	0.96±0.57	3.55±1.76	14.69	1.28	1.24±0.03	**0.85**
5	28, M, 37.9, 127, 183	**0.79**±0	0.83±0.33	4.83±0.89	2.84	1.25	1.10±0.07	1.00
6	21, M, 22.2, 75, 184	0.53±0.47	**0.45**±0.07	2.61±1.61	20.20	1.26	1.10±0.03	0.66
7	32, M, 30.1, 82, 165	1.06±3.86×10^{-3}	1.28±0.19	4.45±2.23	1.83	1.47	1.98±0.09	**0.79**
8	32, M, 27.2, 92, 180	1.08±0.22	1.46±0.89	3.47±2.72	1.65	1.04	1.06±0.04	**0.48**
9	29, M, 19.7, 68, 186	**0.48**±0.40	2.47±0.60	1.34±0.46	0.77	2.12	2.03±0.04	1.26
10	24, F, 22.8, 62, 165	0.34±0.05	**0.24**±1.42×10^{-3}	3.00±1.72	0.53	1.07	1.39±0.03	0.64
11	31, M, 22.5, 68, 174	0.70±5.99×10^{-3}	1.09±9.54×10^{-3}	2.16±1.20	**0.61**	1.66	1.42±0.04	1.05
12	27, M, 24.2, 70, 170	0.83±0.20	0.30±0.10	0.66±0.55	**0.18**	1.28	1.10±0.04	0.94
13	29, M, 23.6, 79, 183	**0.02**±3.49×10^{-4}	0.27±0.37	1.14±0.28	0.09	1.03	1.12±0.04	0.53
14	18, M, 25.6, 83, 180	**1.05**±0.09	1.64±0.37	2.88±1.52	6.52	1.63	2.27±0.04	1.62
15	39, M, 25.0, 74, 172	**0.04**±4.74×10^{-4}	0.14±9.91×10^{-3}	0.45±0.33	0.27	1.45	1.29±0.04	0.89
16	26, M, 21.2, 65, 175	**0.22**±1.02×10^{-5}	**0.22**±0.02	2.09±1.14	6.12	1.25	1.65±0.05	0.50
17	31, M, 28.9, 100, 186	**0.10**±4.53×10^{-3}	**0.10**±3.21×10^{-3}	1.66±1.12	0.49	1.04	1.87±0.19	0.46
18	27, M, 22.3, 70, 177	0.25±3.07×10^{-3}	0.45±0.01	1.64±1.72	**0.20**	1.21	1.07±0.04	0.63
19	30, M, 25.1, 76, 174	0.30±0.07	**0.15**±4.31×10^{-3}	1.18±0.95	0.62	1.45	1.48±0.05	0.71
20	23, M, 24.7, 73, 172	0.69±0.29	**0.56**±0.03	4.43±3.93	0.85	2.35	2.12±0.09	1.42
21	30, F, 19.7, 57, 170	1.04±0.06	1.18±0.37	4.08±1.55	4.29	0.90	1.07±0.14	**0.23**
22	27, M, 27.5, 86, 177	0.22±0.25	**0.10**±6.17×10^{-3}	2.12±0.34	0.38	1.65	1.51±0.07	0.44
23	24, F, 20.8, 60, 170	**0.33**±3.30×10^{-3}	0.81±4.95×10^{-3}	1.76±0.72	7.65	1.44	1.61±0.13	1.12
24	25, M, 27.1, 86, 178	1.91±1.23	**0.32**±0.03	2.21±0.27	7.21	2.74	1.86±0.08	1.14
25	49, M, 23.0, 83, 190	0.72±4.42×10^{-3}	0.20±0.04	0.73±0.75	**0.07**	1.35	1.41±0.04	0.80
26	22, M, 26.0, 92, 188	0.71±0.17	**0.25**±0.02	2.08±1.82	0.49	2.13	2.30±0.26	1.81
27	33, M, 29.2, 82, 167.6	**0.43**±0.33	0.56±0.27	4.57±3.26	0.89	1.45	1.49±0.04	1.00
28	28, M, 21.2, 73, 185.4	0.82±4.21×10^{-3}	**0.62**±0.22	1.64±1.10	3.21	1.72	1.96±0.08	1.38
29	34, M, 25.6, 84, 181	0.32±0	0.47±0.08	1.15±0.54	**0.09**	0.89	0.87±0.06	0.34
30	34, M, 24.4, 64, 162	**0.15** ±4.49×10^{-3}	**0.15**±0.03	3.67±3.14	0.17	1.06	1.28±0.16	0.34
31	32, F, 28.2, 77, 165.1	1.00±1.28	**0.99**±0.12	3.75±1.31	1.47	1.30	2.36±0.13	1.71
32	22, F, 21.0, 51, 156	**0.91**±0.46	1.26±0.57	2.53±1.38	10.23	1.59	2.36±0.04	1.03
33	27, M, 23.8, 77, 180	0.68±0.03	0.71±0.07	0.73±0.23	8.43	1.82	0.98±0.01	**0.54**
34	26, M, 23.7, 83, 187	**0.33** ±0.21	0.61±0.13	2.72±2.27	4.65	1.49	1.55±0.08	0.64
35	28, F, 18.3, 48, 162	0.53 ±0.69	**0.15**±0.05	3.98±1.63	0.43	1.34	1.15±0.03	0.76
36	38, M, 36.9, 120, 180.3	**0.44**±1.13×10^{-3}	1.44±0.31	1.00±0.52	1.68	1.15	1.19±0.16	0.47
37	37, M, 22.1, 68, 175.3	**0.23**±0.07	**0.23**±9.86×10^{-3}	2.42±1.54	0.85	0.74	0.89±0.05	0.40
38	29, M, 25.8, 79, 175	**0.29**±0.08	0.46±0.16	1.37±1.24	6.53	1.13	0.87±0.07	0.47
39	23, M, 23.0, 68, 172	**0.36**±1.75×10^{-3}	0.71±0.38	4.11±3.43	0.52	1.80	1.84±0.08	1.00
40	32, M, 30.9, 99, 179	2.06±0.29	**0.37**±0.04	2.57±2.10	1.94	1.24	1.44±0.08	0.55
	Total average	**0.63**	0.64	2.43	3.02	1.42	1.49	0.83

Table 5.9: The Correlation Coefficients between Performance and Age, Weight, Height, BMI and Ground Truth.

	MI-HE	DL-FUMI	EM-DD	WPPD	CA	EN	HT
Age	-0.12	-0.12	-0.37	-0.28	-0.31	-0.46	-0.35
Weight	0.10	0.04	-0.06	-0.12	-0.04	-0.21	-0.01
Height	-0.09	0.00	-0.35	-0.07	-0.01	-0.11	0.01
BMI	0.15	0.05	0.11	-0.08	-0.04	-0.21	-0.01

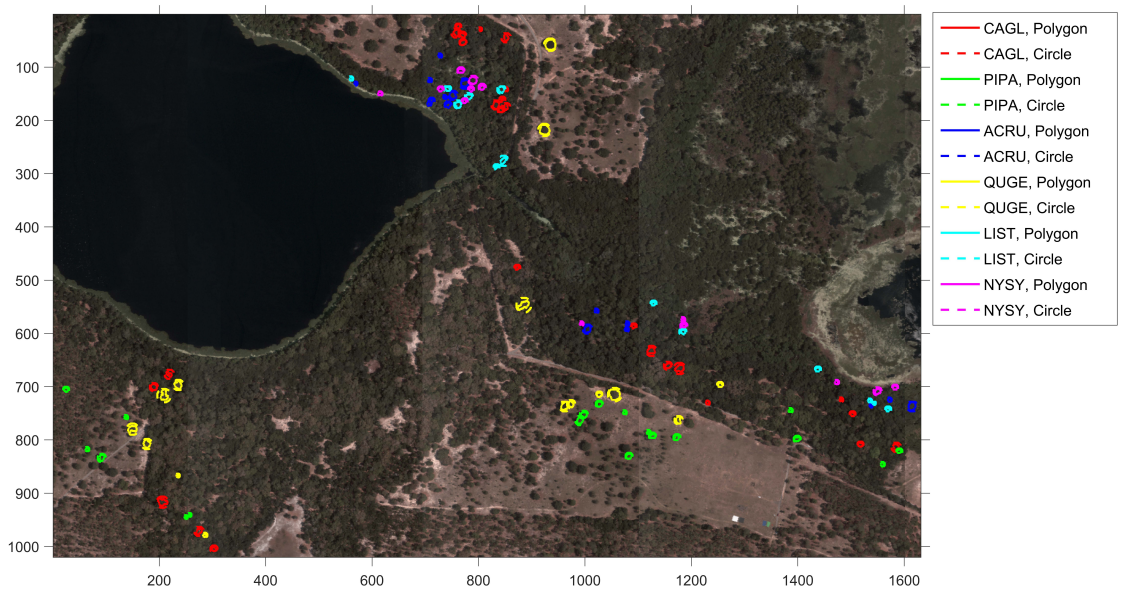
the robustness of the proposed algorithm.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}, \tag{5.1}$$

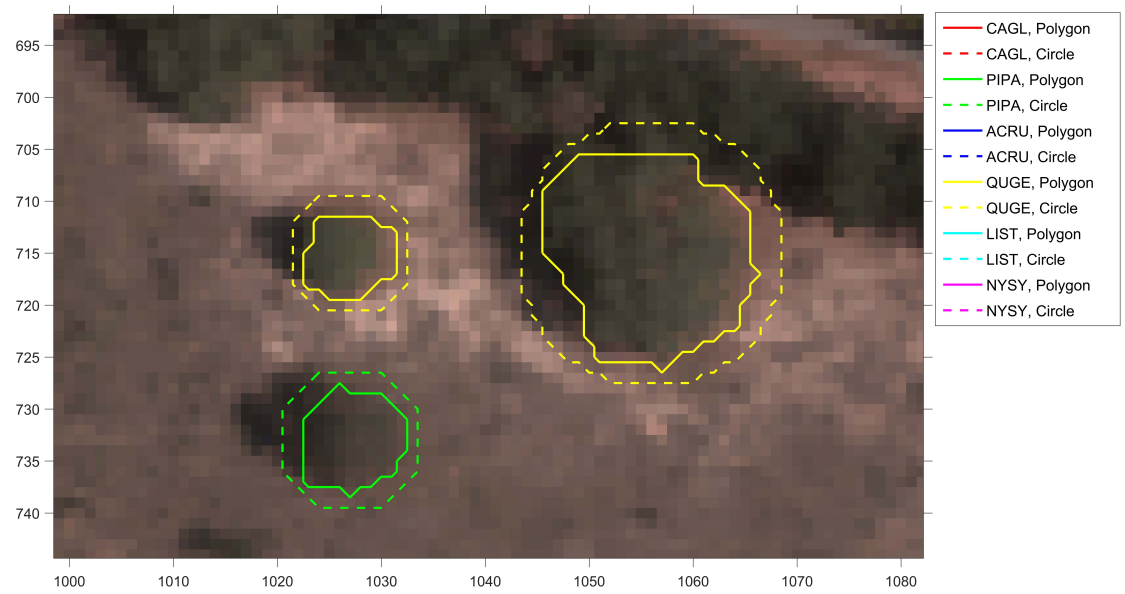
5.4 Tree Species Classification from NEON Data

In this section, we applied MI-HE to a multiple instance tree species classification problem using the subset of National Ecological Observatory Network (NEON) data [132] collected at the Ordway-Swisher Biological Station (OSBS) in north-central Florida, United States. This data contains 1020×1631 pixels with 428 bands corresponding to wavelengths from 380 nm to 2510 nm at a 5 nm spectral sampling interval. The spatial resolution and collection altitude are 1 pixel/m^2 and 1000 meters, respectively.

Fig. 5.29(a) shows the full ground view (RGB) of the hyperspectral data where there are six types of trees named CAGL, PIPA, ACRU, QUGE, LIST and NYSY, denoted by polygons with different colors. The number of labeled trees for each type is: 26, 18, 16, 15, 13 and 13, respectively. Fig. 5.29(b) shows the zoomed view of Fig. 5.29(a), where the solid line denotes the polygon contours of each tree's canopy. These polygons were collected by Stephanie Bohlman and Sarah Graves from the School of Forest Resources and Conservation, University of Florida. The polygon contours were drawn by having an



(a) Full view



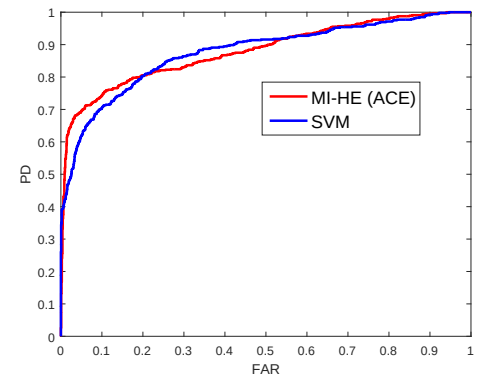
(b) Zoomed view

Figure 5.29: RGB image of NEON OSBS with tree polygons

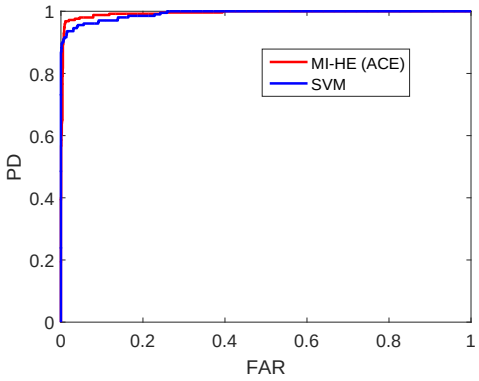
analyst walk under each tree and recording the analyst's trace by a GPS. These polygon contours create an accurately labeled training data for tree species classification. Several existing methods [133, 134] train a SVM from these accurately labeled polygons for tree species classification. However, obtaining accurate tree canopy labels are time consuming, requiring lots of efforts. Furthermore, the canopy labels could be inherently inaccurate, *e.g.*, the accuracy of GPS could drift several meters; the boundary of a tree canopy could be ambiguous.

So here we model this task as a MIL problem and more general, inaccurate labels were generated by drawing a circle including each polygon shown as dashed line in Fig. 5.29(b). The goal is to show the proposed MI-HE is able to do tree species classification well given the inaccurate circle labels. The experiment was conducted by training on randomly selecting 70% of the canopies and testing on the remaining 30% canopies. The testing step was conducted by scoring the testing data point by point, where the testing data were per-pixel labeled according to the polygon data. Hierarchical dimension reduction was applied to reduce the dimensionality of the data to 124. Both the circle data and polygon data were used as training and compared with SVM. The parameter settings of MI-HE for this experiment are $T = 1, M = 8, \rho = 0.5, p = 5, \beta = 1$ and $\lambda = 1 \times 10^{-3}$. The experiments were repeated for five times and the median performance (AUC) was shown.

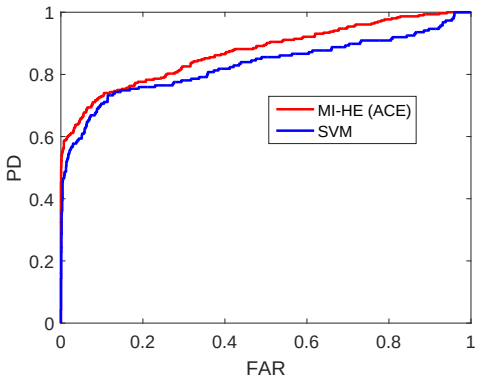
Fig. 5.30 and 5.31 show the ROC curves of MI-HE and SVM that train and test on polygon data and circle data, respectively. Tab. 5.10 shows the detailed AUCs for each run. For training and testing on the polygon data, SVM outperforms MI-HE only on tree type QUGE and has close overall performance to MI-HE. However, for training and testing on the circle data, MI-HE outperforms SVM on each type of the classification. Furthermore, compared with the results from the polygon data, although MI-HE and SVM both provide



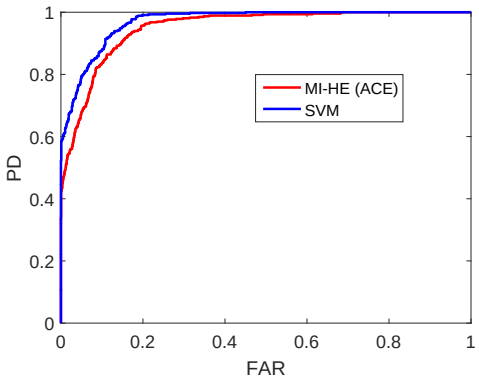
(a) CAGL



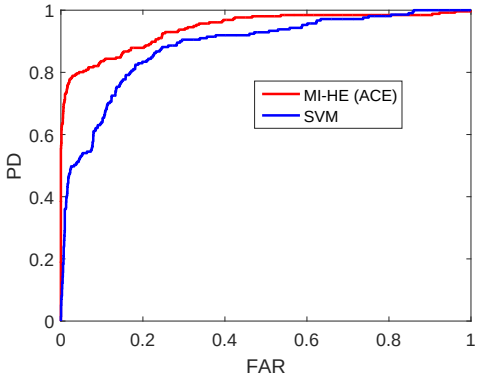
(b) PIPA



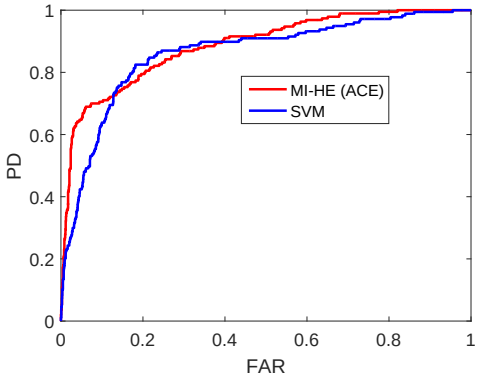
(c) ACRU



(d) QUGE

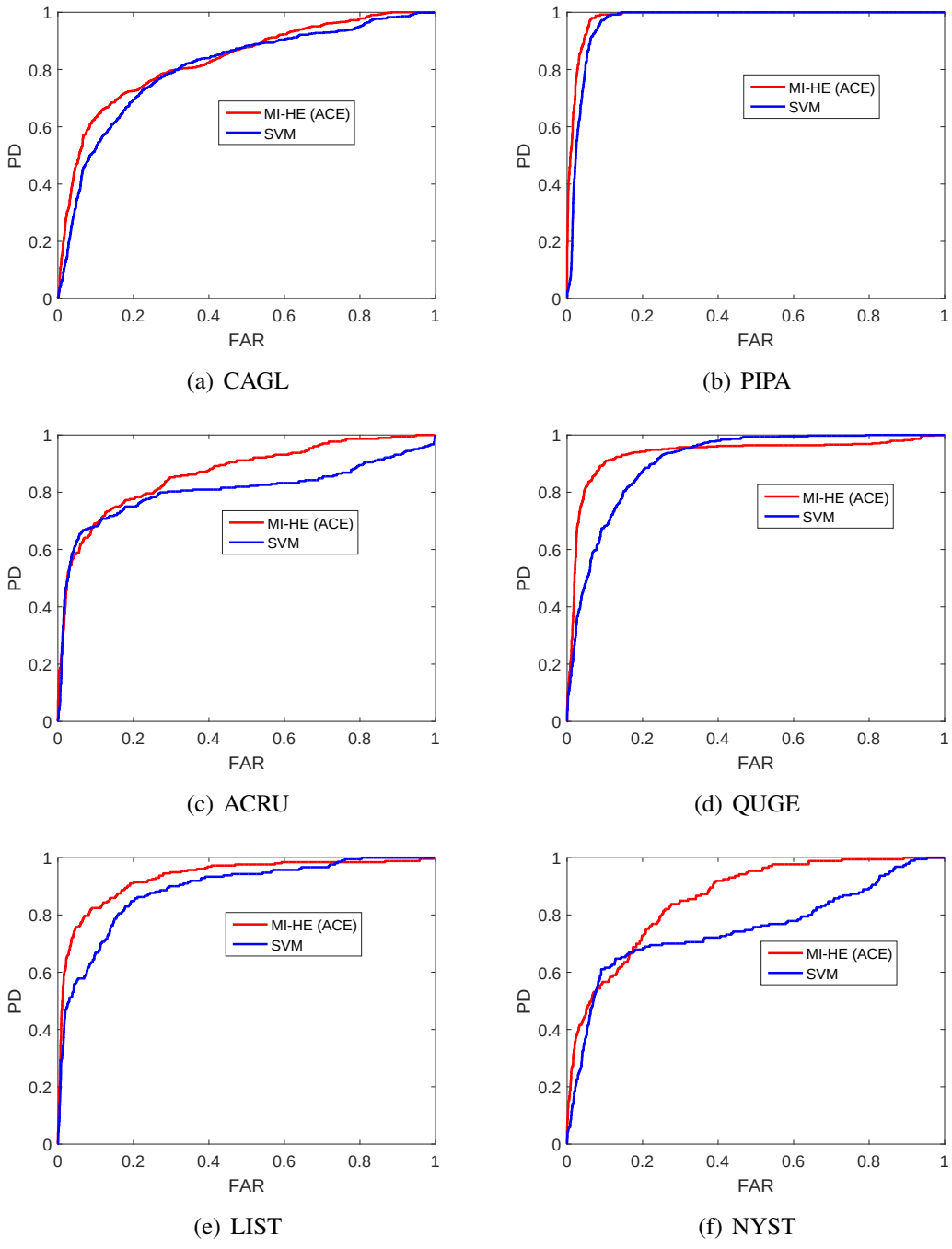


(e) LIST



(f) NYST

Figure 5.30: ROC curves of MI-HE and SVM on polygon data



(a) CAGL (b) PIPA

(c) ACRU (d) QUGE

(e) LIST (f) NYST

Figure 5.31: ROC curves of MI-HE and SVM on circle data

Table 5.10: Tree Species Classification Results (AUCs), Bold for the Best, Underline for the Second Best

Species	Polygon Data		Circle Data	
	MI-HE	SVM	MI-HE	SVM
CAGL	**0.880**	0.876	**0.833**	0.807
PIPA	**0.994**	0.992	**0.984**	0.969
ACRU	**0.873**	0.829	**0.867**	0.806
QUGE	0.950	**0.970**	**0.936**	0.910
LIST	**0.942**	0.884	**0.933**	0.890
NYSY	**0.887**	0.862	**0.861**	0.755
Average	**0.921**	0.902	**0.902**	0.856

decreased performance on the circle data, SVM suffers more from the label uncertainty given the circle data.

Chapter 6

Conclusion and Future Work

In this work the MI-HE target concept learning framework for MIL problems is proposed and investigated. MI-HE is able to learn multiple discriminative target concepts from ambiguously labeled data. After learning target concepts, target detection can be conducted by applying the estimated target concept to any signature based detector. Comprehensive experiments show that proposed MI-HE is effective in learning discriminative target concept and achieves superior performance over comparison algorithms in several scenarios.

Future work will include developing an automatic parameter setting tool for MI-HE. For example, the number of background concepts could be determined by examining cluster validity measures from some efficient unsupervised learning methods (*e.g.*, K-means [124], VCA [109], PM-LDA [135–137], K-SVD [128]); the sparsity level λ could be empirically estimated from the magnitude of the training data. Furthermore, improving the computational efficiency of MI-HE is also important. For example, the sparse unmixing step in MI-HE can be executed in parallel using many-core devices and accelerators (e.g. GPUs, Intel Xeon Phi and FPGA [138–140]); instead of using gradient descent, conju-

gate gradient can be adopted for dictionary learning. Finally, more experiments will be conducted. We expect to get more tree labeling information for the NEON data and conduct comprehensive experiments with more tree species. We also plan to apply MI-HE to the long-term monitored data acquired at TigerPlace, an active aging-in-place retirement community developed by the MU Sinclair School of Nursing and CERT at the University of Missouri, and study about how the position, posture, and body movement affect the performance of the proposed algorithm.

Bibliography

[1] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[2] O. Maron and T. Lozano-Perez, "A framework for multiple-instance learning," *Advances in Neural Information Processing Systems (NIPS)*, vol. 10, pp. 570–576, 1998.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[4] D. Manolakis, D. Marden, and G. A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 79–116, 2003.

[5] N. M. Nasrabadi, "Hyperspectral target detection: An overview of current and future challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 34–44, 2014.

[6] I. Starr, A. Rawson, H. Schroeder, and N. Joseph, "Studies on the estimation of car-

diac ouptut in man, and of abnormalities in cardiac function, from the heart's recoil and the blood's impacts; the ballistocardiogram," *American Journal of Physiology–Legacy Content*, vol. 127, no. 1, pp. 1–28, 1939.

[7] E. Pinheiro, O. Postolache, and P. Girão, "Theory and developments in an unobtrusive cardiovascular system representation: Ballistocardiography," *The Open Biomedical Engineering Journal*, vol. 4, p. 201, 2010.

[8] O. T. Inan, P.-F. Migeotte, K.-S. Park, M. Etemadi, K. Tavakolian, R. Casanella, J. Zanetti, J. Tank, I. Funtova, G. K. Prisk *et al.*, "Ballistocardiography and seismocardiography: A review of recent advances," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1414–1427, 2015.

[9] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 17–28, 2002.

[10] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.

[11] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[12] S. E. Yuksel, J. Bolton, and P. Gader, "Multiple-instance hidden markov models with applications to landmine detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 12, pp. 6766–6775, 2015.

[13] A. Zare, J. Bolton, P. Gader, and M. Schatten, "Vegetation mapping for landmine detection using long-wave hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 1, pp. 172–178, Jan 2008.

[14] G. Mahajan, R. Sahoo, R. Pandey, V. Gupta, and D. Kumar, "Using hyperspectral remote sensing techniques to monitor nitrogen, phosphorus, sulphur and potassium in wheat (triticum aestivum l.)," *Precision Agriculture*, vol. 15, no. 5, pp. 499–522, 2014.

[15] Z. Wang, L. Lan, and S. Vucetic, "Mixture model for multiple instance regression and applications in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2226–2237, 2012.

[16] R. Pike, G. Lu, D. Wang, Z. G. Chen, and B. Fei, "A minimum spanning forest-based method for noninvasive cancer detection with hyperspectral imaging," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 653–663, 2016.

[17] A. Pardo, E. Real, V. Krishnaswamy, J. M. López-Higuera, B. W. Pogue, and O. M. Conde, "Directional kernel density estimation for classification of breast tissue spectra," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 64–73, 2017.

[18] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated hyperspectral cueing for civilian search and rescue," *Proceedings of the IEEE*, vol. 97, no. 6, pp. 1031–1055, 2009.

[19] M. Lara, L. Lleó, B. Diezma-Iglesias, J.-M. Roger, and M. Ruiz-Altisent, "Monitoring spinach shelf-life with hyperspectral image through packaging films," *Journal of Food Engineering*, vol. 119, no. 2, pp. 353–361, 2013.

[20] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 44, no. 2-3, pp. 127–143, 1993.

[21] J. M. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, Apr. 2012.

[22] N. Keshava, J. Kerekes, D. Manolakis, and G. Shaw, "An algorithm taxonomy for hyperspectral unmixing," *Proceedings of the SPIE*, vol. 4049, pp. 42–63, 2000.

[23] N. Keshava, "A survey of spectral unmixing algorithms," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 55–78, 2003.

[24] M. Parente and A. Plaza, "Survey of geometric and statistical unmixing algorithms for hyperspectral images," in *2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2010, pp. 1–4.

[25] J. M. Bioucas-Dias and A. Plaza, "An overview on hyperspectral unmixing: Geometrical, statistical, and sparse regression based approaches," *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1135–1138, 2011.

[26] T.-H. Chan, W.K.-Ma, A. Ambikapathi, and C.-Y. Chi, "A simplex volume maximization framework for hyperspectral endmember extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4177–4193, June 2011.

[27] J. Wang and C.-I. Chang, "Applications of independent component analysis in endmember extraction and abundance quantification for hyperspectral imagery," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 44, no. 9, pp. 2601–2616, Sep. 2006.

[28] C.-I. Chang, C.-C. Wu, C.-S. Lo, and M.-L. Chang, "Real-time simplex growing algorithms for hyperspectral endmember extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 4, pp. 1834–1850, April 2010.

[29] M.-D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 3, pp. 542–552, May 1994.

[30] A. Ifarraguerri and C.-I. Chang, "Multispectral and hyperspectral image analysis with convex cones," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 73, no. 2, pp. 756–770, Mar. 1999.

[31] J. M. P. Nascimento and J. M. Bioucas Dias, "Does independent component analysis play a role in unmixing hyperspectral data?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 1, pp. 175–187, 2005.

[32] M. Berman *et al.*, "ICE: A statistical approach to identifying endmembers in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, pp. 2085–2095, Oct. 2004.

[33] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 161–173, Jan. 2009.

[34] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, Mar. 2007.

[35] A. Zare and P. Gader, "Sparsity promoting iterated constrained endmember detection for hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 3, pp. 446–450, July 2007.

[36] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2014–2039, June 2011.

[37] Y. Zhong, R. Feng, and L. Zhang, "Non-local sparse unmixing for hyperspectral remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 1889–1909, June 2014.

[38] X. Lu *et al.*, "Manifold regularized sparse NMF for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2815–2826, 2013.

[39] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4484–4502, 2012.

[40] F. Chen and Y. Zhang, "Sparse hyperspectral unmixing based on constrained $\ell_p - \ell_2$ optimization," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1142–1146, 2013.

[41] Z. Shi, W. Tang, Z. Duren, and Z. Jiang, "Subspace matching pursuit for sparse unmixing of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 6, pp. 3256–3274, June 2014.

[42] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "Spatial/spectral endmember extraction by multidimensional morphological operators," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 9, pp. 2025–2041, Sep. 2002.

[43] D. M. Rogge, B. Rivard, J. Zhang, A. Sanchez, J. Harris, and J. Feng, "Integration of spatial-spectral information for the improved extraction of endmembers," *Remote Sensing of Environment*, vol. 110, pp. 287–303, 2007.

[44] A. Zare, O. Bchir, H. Frigui, and P. Gader, "Spatially-smooth piece-wise convex endmember detection," in *2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2010, pp. 1–4.

[45] A. Zare and P. Gader, "Piece-wise convex spatial-spectral unmixing of hyperspectral imagery using possibilistic and fuzzy clustering," in *IEEE International Conference on Fuzzy Systems*, 2011, pp. 741–746.

[46] M. Xu, B. Du, and L. Zhang, "Spatial-spectral information based abundance-constrained endmember extraction methods," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 1939–1404, June 2014.

[47] J. Chen, C. Richard, and P. Honeine, "Nonlinear unmixing of hyperspectral data based on a linear-mixture/nonlinear-fluctuation model," *IEEE Transactions on Signal Processing*, vol. 61, no. 2, pp. 480–492, 2013.

[48] Y. Altmann, M. Pereyra, and S. McLaughlin, "Bayesian nonlinear hyperspectral unmixing with spatial residual component analysis," *IEEE Transactions on Computational Imaging*, vol. 1, no. 3, pp. 174–185, 2015.

[49] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

[50] D. Manolakis, E. Truslow, M. Pieper, T. Cooley, and M. Brueggeman, "Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 24–33, 2014.

[51] P. Gader, A. Zare *et al.*, "MUUFL gulfport hyperspectral and lidar airborne data set," University of Florida, Gainesville, FL, REP-2013-570, Tech. Rep., Oct. 2013.

[52] M. Skubic, R. D. Guevara, and M. Rantz, "Automated health alerts using in-home sensor data for embedded health assessment," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 3, pp. 1–11, 2015.

[53] L. Rosales, M. Skubic, D. Heise, M. J. Devaney, and M. Schaumburg, "Heartbeat detection from a hydraulic bed sensor using a clustering approach," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2012, pp. 2383–2387.

[54] D. Heise, L. Rosales, M. Sheahen, B.-Y. Su, and M. Skubic, "Non-invasive measurement of heartbeat with a hydraulic bed sensor progress, challenges, and opportunities," in *International Instrumentation and Measurement Technology Conference*. IEEE, 2013, pp. 397–402.

[55] L. Rosales, B. Y. Su, M. Skubic, and K. C. Ho, "Heart rate monitoring using hydraulic bedsensor ballistocardiogram," *Journal of Ambient Intelligence and Smart Environment*, vol. 9, no. 2, pp. 193–207, Feb. 2017.

[56] J. Broadwater and R. Chellappa, "Hybrid detectors for subpixel targets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1891–1903, Nov. 2007.

[57] M. T. Eismann, *Hyperspectral Remote Sensing*. SPIE Press, 2012.

[58] S. M. Kay, *Fundamental of Statistical Signal Processing: Volume II – Detection Theory*. Prentice-Hall, 1993.

[59] S. Matteoli, M. Diani, and J. Theiler, "An overview of background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2317–2336, June 2014.

[60] J. Theiler and B. R. Foy, "Effect of signal contamination in matched-filter detection of the signal on a cluttered background," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 98–102, Jan 2006.

[61] N. M. Nasrabadi, "Regularized spectral matched filter for target recognition in hyperspectral imagery," *IEEE Signal Processing Letters*, vol. 15, pp. 317–320, 2008.

[62] S. Kraut and L. Scharf, "The CFAR adaptive subspace detector is a scale-invariant GLRT," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2538 –2541, Sept. 1999.

[63] S. Kraut, L. Scharf, and L. McWhorter, "Adaptive subspace detectors," *IEEE Transactions on Signal Processing*, vol. 49, no. 1, pp. 1–16, 2001.

[64] W. F. Basener, "Clutter and anomaly removal for enhanced target detection," in *Proceedings of the SPIE*, vol. 7695, 2010, p. 769525.

[65] J. Broadwater, R. Meth, and R. Chellappa, "A hybrid algorithm for subpixel detection in hyperspectral imagery," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 3. IEEE, 2004, pp. 1601–1604.

[66] D. C. Heinz *et al.*, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529–545, 2001.

[67] A. Zare and C. Jiao, "Extended functions of multiple instances for target characterization," in *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2014, pp. 1–4.

[68] ——, "Functions of multiple instances for sub-pixel target characterization in hyperspectral imagery," in *SPIE Defense + Security*. International Society for Optics and Photonics, 2015, pp. 947 212–947 212.

[69] C. Jiao and A. Zare, "Functions of multiple instances for learning target signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4670 – 4686, 2015.

[70] C. Jiao, P. Lyons, A. Zare, L. Rosales, and M. Skubic, "Heart beat characterization from ballistocardiogram signals using extended functions of multiple instances," in *38th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2016, pp. 756–760.

[71] C. Jiao and A. Zare, "Multiple instance dictionary learning using functions of multiple instances," in *International Conference on Pattern Recognition (ICPR)*, December 2016, pp. 2688–2693.

[72] C. Jiao, B.-Y. Su, P. Lyons, A. Zare, K. Ho, and M. Skubic, "Multiple instance dictionary learning for beat-to-beat heart rate monitoring from ballistocardiograms," *arXiv preprint arXiv:1706.03373*, 2017.

[73] A. Zare, C. Jiao, and T. Glenn, "Discriminative multiple instance hyperspectral target characterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, In press.

[74] C. Jiao and A. Zare, "Multiple instance hybrid estimator for learning target signatures," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 1–4.

[75] C. Jiao, A. Zare, and R. G. McGarvey, "Multiple instance hybrid estimator for hyperspectral target characterization and sub-pixel target detection," *arXiv preprint arXiv:1710.11599*, 2017.

[76] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American statistical Association*, vol. 76, no. 376, pp. 817–823, 1981.

[77] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification." in *International Conference on Machine Learning*, vol. 98, 1998, pp. 341–349.

[78] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," *Advances in Neural Information Processing Systems (NIPS)*, vol. 2, pp. 1073–1080, 2002.

[79] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical recipes in C: the art of scientific programming*. Cambridge University Press, 1992.

[80] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2002, pp. 561–568.

[81] A. Shrivastava, J. K. Pillai, V. M. Patel, and R. Chellappa, "Dictionary-based multiple instance learning," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 160–164.

[82] A. Shrivastava, V. M. Patel, j. K. Pillai, and R. Chellappa, "Generalized dictionaries for multiple instance learning," *International Journal of Computer Vision*, vol. 114, no. 2, pp. 288–305, Septmber 2015.

[83] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[84] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[85] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[86] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[87] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013.

[88] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.

[89] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[90] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[91] Y. Chen, J. Bi, and J. Z. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[92] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," *Advances in Neural Information Processing Systems (NIPS)*, vol. 16, no. 1, pp. 49–56, 2004.

[93] K. Bennett and O. Mangasarian, "Combining support vector and mathematical programming methods for induction," *Advances in Kernel Methods-SV Learning*, pp. 307–326, 1999.

[94] A. Smola, B. Scholkopf, and G. Ratsch, "Linear programs for automatic accuracy control in regression," in *9th International Conference on Artificial Neural Networks*, vol. 2. IET, 1999, pp. 575–580.

[95] J. Huo, Y. Gao, W. Yang, and H. Yin, "Abnormal event detection via multi-instance dictionary learning," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2012, pp. 76–83.

[96] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[97] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning," in *International Conference on Machine Learning*, 2013, pp. 846–854.

[98] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[99] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[100] Z. Fu, A. Robles-Kelly, and J. Zhou, "Milis: Multiple instance learning with instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 958–977, 2011.

[101] Z.-H. Zhou and J.-M. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *Proceedings of the 24th International Conference on Machine Learning*. ACM, 2007, pp. 1167–1174.

[102] Q. Wang, L. Ruan, and L. Si, "Adaptive knowledge transfer for multiple instance learning in image classification," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 1334–1340.

[103] K. Ali and K. Saenko, "Confidence-rated multiple instance boosting for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 2433–2440.

[104] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko, "Detector discovery in the wild: Joint multiple instance and representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp. 2883–2891.

[105] W. Li and N. Vasconcelos, "Multiple instance learning for soft bags via top instances," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp. 4277–4285.

[106] M. Rastegari, H. Hajishirzi, and A. Farhadi, "Discriminative and consistent similarities in instance-level multiple instance learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp. 740–748.

[107] A. Zare and P. Gader, "Pattern recognition using functions of multiple instances," *International Conference on Pattern Recognition (ICPR)*, pp. 1092–1095, Aug. 2010.

[108] A. Zare, P. Gader, J. Bolton, S. Yuksel, T. Dubroca, and R. Close, "Sub-pixel target spectra estimation using functions of multiple instances," in *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2011, pp. 1–4.

[109] J. M. Nascimento and J. M. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.

[110] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3501–3508.

[111] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[112] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[113] M. A. Figueiredo and R. D. Nowak, "An em algorithm for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 906–916, 2003.

[114] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, pp. 1413–1457, 2004.

[115] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[116] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[117] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[118] S. Mallat, *A wavelet tour of signal processing, Third Edition: The Sparse Way*.

[119] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[120] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.

[121] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *International Conference on Computer Vision*. IEEE, 2011, pp. 543–550.

[122] ——, "Sparse representation based fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.

[123] X. Li, R. Guo, and C. Chen, "Robust pedestrian tracking and recognition from FLIR video: A unified approach via sparse coding," *Sensors*, vol. 14, no. 6, pp. 11 245–11 259, 2014.

[124] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[125] A. Baldridge, S. Hook, C. Grove, and G. Rivera, "The ASTER spectral library version 2.0," *Remote Sensing of Environment*, vol. 113, no. 4, pp. 711–715, 2009.

[126] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[127] T. Glenn, A. Zare, P. Gader, and D. Dranishnikov, "Bullwinkle: Scoring code for sub-pixel targets (version 1.0) [software]," 2013, http://engineers.missouri.edu/zarea/code/.

[128] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[129] D. Heise and M. Skubic, "Monitoring pulse and respiration with a non-invasive hydraulic bed sensor," in *32th International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010, pp. 2119–2123.

[130] K. Lydon, B. Y. Su, L. Rosales, M. Enayati, K. C. Ho, M. Rantz, and M. Skubic, "Robust heartbeat detection from in-home ballistocardiogram signals of older adults using a bed sensor," in *37th International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015, pp. 7175–7179.

[131] B. Y. Su, K. C. Ho, M. Skubic, and L. Rosales, "Pulse rate estimation using hydraulic bed sensor," in *34th International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 2587–2590.

[132] National Ecological Observatory Network, 2016, Data accessed on Jan., 2016. Available on-line http://data.neonscience.org/ from Battelle, Boulder, CO, USA.

[133] S. J. Graves, G. P. Asner, R. E. Martin, C. B. Anderson, M. S. Colgan, L. Kalantari, and S. A. Bohlman, "Tree species abundance predictions in a tropical agricultural landscape with a supervised classification model and imbalanced data," *Remote Sensing*, vol. 8, no. 2, p. 161, 2016.

[134] M. S. Nia, D. Z. Wang, S. A. Bohlman, P. Gader, S. J. Graves, and M. Petrovic, "Impact of atmospheric correction and image filtering on hyperspectral classification of tree species using support vector machine," *Journal of Applied Remote Sensing*, vol. 9, no. 1, pp. 095 990–095 990, 2015.

[135] C. Chen, A. Zare, and J. T. Cobb, "Partial membership latent dirichlet allocation for image segmentation," in *International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2368–2373.

[136] C. Chen, "Partial membership latent dirichlet allocation," *Ph.D. Thesis*, May 2016.

[137] C. Chen, A. Zare, H. N. Trinh *et al.*, "Partial membership latent dirichlet allocation for soft image segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5590–5602, 2017.

[138] D. Li, K. Sajjapongse, H. Truong *et al.*, "A distributed CPU-GPU framework for pairwise alignments on large-scale sequence datasets," in *IEEE 24th International Conference on Application-Specific Systems, Architectures and Processors (ASAP)*, 2013, pp. 329–338.

[139] D. Li, S. Chakradhar, and M. Becchi, "Grapid: A compilation and runtime framework for rapid prototyping of graph applications on many-core processors," in *20th IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, 2014, pp. 174–182.

[140] T. Song, D. Li, and Y. Yao, "Multi-source data oriented flexible real-time information fusion platform on FPGA," in *International Conference on Electronics, Communications and Control (ICECC)*, 2011, pp. 4401–4404.

VITA

Changzhe Jiao was born in April 1985, in Xi'an city, China. He received the Bachelor and Master degrees in Automation and Control Theory from Xidian University in July 2007 and June 2012, respectively. He joined the Department of Electrical and Computer Engineering at University of Kentucky as a Ph.D. student in August 2012 and transferred to the Department of Electrical and Computer Engineering at the University of Missouri in January 2013. His research interests include pattern recognition, multiple instance learning and hyperspectral image analysis.